VIRGINIA TEACHER EVALUATION AND VIRGINIA PERFORMANCE-PAY INCENTIVES (VPPI) PILOT

An Evaluation Report

James Stronge
Thomas Ward
Xianxuan Xu

The College of William and Mary

February 2013

Contents

Introduction

Teacher evaluation matters because teaching matters. In fact, "the core of education *is* teaching and learning, and the teaching-learning connection works best when we have effective teachers working with every student every day."[1] The quality of an education system cannot exceed the quality of its teachers.[2] Teacher effectiveness is the most influential school-related factor in student achievement. During the past decade, there has been a growing interest in better understanding what constitutes teacher effectiveness. This focus has presented challenges as well as opportunities for the policymakers, researchers, and practitioners to develop a teacher evaluation system that can efficiently and reliably measure teacher performance.

The role of a teacher requires a performance evaluation system that acknowledges the complexities of the job. Teachers have a challenging task in meeting the educational needs of an educationally diverse student population, and good evaluation is necessary to provide the teachers with the support, recognition, and guidance they need to sustain and improve their efforts.[3] Traditional teacher evaluation is inadequate both for differentiating between more and less effective teachers and as a basis for guiding improvements in teaching skills.[4] Traditional teacher evaluation is usually a process that typically results in universally high ratings for all teachers. It is impossible to acknowledge and act on differences in teacher performance.[5] In this era of accountability, especially with the federal initiatives and the Elementary and Secondary Education Act (ESEA) Flexibility Plan of the federal government, more and more states are starting to put emphasis on the rigor and usefulness of teacher evaluation and its potential as an information tool to measure teacher performance, recognize effective teaching, identify teacher needs, and promote teacher growth. The new teacher evaluation system envisioned by the Commonwealth of Virginia was based on a comprehensive conception of professional expectations for teachers, as indicated by the standards, sample indicators, and performance rubrics that were derived from research and best practices. Also incorporated were multiple sources of information to demonstrate a complete portrait of teacher performance. Furthermore, variations in teacher performance will be recorded and used meaningfully for teacher professional development.

The purpose of this report is to evaluate the validity of the 2011-2012 pilot of this teacher evaluation system implemented in 25 schools. The Virginia Performance-Pay Incentives (VPPI) pilot provided funding to award competitive grants to Hard-to-Staff (HTS) schools in school divisions throughout Virginia. Nine HTS schools, representing six school divisions, were selected to participate in the pilot through the competitive grant process. In an effort to increase the number of pilot schools, participation in the pilot was expanded to include schools that received federal School Improvement Grants (SIG). Sixteen SIG schools, representing eight school divisions, participated in the pilot.

Pilot Process: An Introduction to the Evaluation System

In July through December 2010, the Virginia Department of Education established a work group that involved diverse stakeholders for the purpose of developing new teacher performance standards and an accompanying performance evaluation system to be implemented statewide. The concerted work of this group resulted in the revision of selected existing documents, and the development of new teacher performance standards and a comprehensive teacher evaluation system. In April 2011, the Virginia Board of Education adopted the new *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*, defining the criteria expected when teachers perform their major duties. Pursuant to state law, teacher evaluations must be consistent with the performance standards (objectives) included in the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*, effective July 1, 2012, and school boards' procedures for evaluating instructional personnel must address student academic progress.

> ***§ 22.1-253.13:5. Standard 5. Quality of classroom instruction and educational leadership****.*
>
> *B. Consistent with the finding that leadership is essential for the advancement of public education in the Commonwealth, teacher, administrator, and superintendent evaluations* ***shall be consistent with the performance objectives included in the Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers****, Administrators, and Superintendents. Teacher evaluations shall include regular observation and evidence that instruction is aligned with the school's curriculum. Evaluations shall include identification of areas of individual strengths and weaknesses and recommendations for appropriate professional activities.* [***emphasis added***]
>
> ***§ 22.1-295 Employment of teachers.***
>
> *C. School Boards shall develop a procedure for use by division superintendents and principals in evaluating instructional personnel that is appropriate to the tasks performed and addresses, among other things,* ***student academic progress*** *and the skills and knowledge of instructional personnel, including, but not limited to, instructional methodology, classroom management, and subject matter knowledge.* [***emphasis added***]

The Board established guidelines that provide school divisions with a model evaluation system. Properly implemented, the evaluation system could provide school divisions with the information needed to support teacher professional growth and to guide personnel decisions, including providing for differentiated compensation or performance-based pay.

The purposes of this new teacher evaluation program are threefold:

1. To improve the effectiveness of teacher performance evaluation instruments and procedures, specifically to include measures of student performance as part of the evaluation process;
2. To apply performance evaluation in meaningful ways to support and improve the performance of the teachers; and

3. To initiate new performance pay methods for teachers.

The key features of the new teacher evaluation system includ:

1. Build on clearly defined job duties. Performance evaluation needs to be built upon clear and reasonable duties of the teachers – "evaluate teachers on what they are hired to do."[6] The new teacher evaluation system was established on the basis of an explicit and accurate description of the work of teachers. It sets forth seven performance standards for all Virginia teachers. Pursuant to state law, teacher evaluations must be consistent with the following performance standards:

Performance Standard 1: Professional Knowledge
The teacher demonstrates an understanding of the curriculum, subject content, and the developmental needs of students by providing relevant learning experiences.

Performance Standard 2: Instructional Planning
The teacher plans using the Virginia Standards of Learning, the school's curriculum, effective strategies, resources, and data to meet the needs of all students.

Performance Standard 3: Instructional Delivery
The teacher effectively engages students in learning by using a variety of instructional strategies in order to meet individual learning needs.

Performance Standard 4: Assessment of and for Student Learning
The teacher systematically gathers, analyzes, and uses all relevant data to measure student academic progress, guide instructional content and delivery methods, and provide timely feedback to both students and parents throughout the school year.

Performance Standard 5: Learning Environment
The teacher uses resources, routines, and procedures to provide a respectful, positive, safe, student-centered environment that is conducive to learning.

Performance Standard 6: Professionalism
The teacher maintains a commitment to professional ethics, communicates effectively, and takes responsibility for and participates in professional growth that results in enhanced student learning.

Performance Standard 7: Student Academic Progress
The work of the teacher results in acceptable, measurable, and appropriate student academic progress.

These seven standards represent the broad domains of a teacher's practice and provide explicit performance expectations. The standards were derived from research and theory on teaching, and they are consensus based. Each performance standard contains performance indicators that identify the key activities that effective teachers demonstrate as they fulfill the work of the performance standards.

2. Evaluate teacher's skills and behaviors that have a direct impact on learning outcomes. Each of seven standards is realistic and research-informed. In addition, these standards include both the processes and the results (i.e., student academic progress) of teaching. The research base behind each of the standards represents a close connection between teacher effectiveness research and teacher evaluation.

3. Use rubrics to rate teacher performance on each standard as defined by a behaviorally-anchored rating scale, which includes a description of performance expected at each level of "exemplary," "proficient," "developing/needs improvement," and "unacceptable."

4. Use multiple data sources. Multiple information sources are required or recommended to be included to help document more comprehensively the performance of teachers. They includ observation, student academic progress, portfolios/documentation log, student survey, and self-evaluation.

5. Use the collected data to inform personnel decisions. Evaluation is a tool, not the outcome. It serves as a systematic tool that enables data-driven personnel and student improvement decisions. If the supervisor and teacher have carefully designed ways to obtain feedback on specific job duties, there should be ample information to help make a well-founded and objective evaluation. This more comprehensive approach to evaluation should provide a strong foundation upon which to make personnel decisions regarding granting tenure, promotion, professional development, compensation, and dismissal.

Pilot Sites

In the 2011-2012 pilot, performance-pay pilot schools were required to use the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers.* Teachers in 25 schools representing 13 schools divisions participated in the Virginia Performance-Pay Incentives (VPPI) pilot. Nine schools were funded through the state funding for HTS schools, and 16 were funded through the federal School Improvement Grants (SIG). The HTS pilot schools included nine schools in six school divisions:

Accomack County Public Schools
Kegotank Elementary School
Pungoteague Elementary School
Caroline County Public Schools
Caroline High School
Madison Elementary School
Dinwiddie County Public Schools
Dinwiddie County Middle School
Greensville County Public Schools
Edward W. Wyatt Middle School
Patrick County Public Schools
Blue Ridge Elementary School
Hardin Reynolds Memorial School

Roanoke City Public Schools
Patrick Henry High School

The 2011 Virginia General Assembly approved Governor McDonnell's request for funding to reward teachers in Hard-to-Staff schools based on student academic progress and other performance measures during the 2011-2012 school year. The program authorized incentive payments of up to $5,000 for teachers earning exemplary ratings. The participating pilot schools applying for state funds must have met at least four of eight HTS criteria associated with schools that may have difficulty recruiting and retaining effective teachers. The criteria were:

- Accredited with warning;
- Average daily attendance rate was two percentage points below the statewide average;
- Percent of special education students exceeded 150 percent of the statewide average;
- Percent of limited English Proficient (LEP) students exceeded 150 percent of the statewide average;
- Percent of teachers with provisional licenses exceeded 150 percent of the statewide average;
- Percent of special education teachers with provisional special education licenses exceeded 150 percent of the statewide average;
- Percent of inexperienced teachers (0 years of teaching experience) hired to total teachers exceeded 150 percent of the statewide percentage; and
- School had one or more inexperienced teachers (0 years of teaching experience) in a critical shortage area.

The SIG pilot schools included 16 schools in eight school divisions:

Colonial Beach Public Schools
Colonial Beach High School
Fluvanna County Public Schools
Central Elementary School
Columbia Elementary School
Cunningham Elementary School
Franklin City Public Schools
Franklin High School
Hopewell City Public Schools
Hopewell High School
Northampton County Public Schools
Kiptopeke Elementary School
Northampton High School
Petersburg City Public Schools
A.P. Hill Elementary School
Peabody Middle School
Richmond City Public Schools
Armstrong High School
Boushall Middle School
Fred Thompson Middle School

Roanoke City Public Schools

Fleming High School
Lincoln Terrace Elementary School
Westside Elementary School

The pilot schools/divisions developed different policies regarding their level of participation. Some pilot schools mandated certain subjects/grade levels or all teachers would participate in the pilot, while others made teacher participation optional. In order to examine the level of participation in each school, the principals and the division contact person were requested to respond to the following questions:

- Were teachers at our school mandated to participate in the performance-pay pilot?
- If "Yes", were all teachers mandated to participate or just teachers of certain subjects/grade levels?
- If "No", how many of the teachers declined participation while they had an option?

Table 1 provides a summary of the level of participation in each of the HTS pilot schools, as reported by school principals or the division contact person:

Table 1. Level of Participation: HTS Schools

Accomack County Public Schools	
Kegotank Elementary School	The school division selected who would participate and what grade levels. There was not an opt-out option. Teachers of grades 2-5 were selected for the school.
Pungoteague Elementary School	The school division selected who would participate and what grade levels. There was not an opt-out option. Teachers of grades 2-5 were selected for the school.
Caroline County Public Schools	
Caroline High School	Teachers were not mandated to participate, and 10 out of 67 teachers declined participation.
Madison Elementary School	Participation was not mandated, but all of the 51 teachers participated.
Dinwiddie County Public Schools	
Dinwiddie County Middle School	All teachers were required to be evaluated with the new teacher evaluation system to avoid having two different evaluation processes in the same building, but the actual teachers included in the VPPI pilot were teachers of English, mathematics, history, and social science, and a few teachers of special education.

Greensville County Public Schools	
Edward W. Wyatt Middle School	Participation was voluntary for teachers whose students took a Standards of Learning (SOL) test. Twenty-four (24) teachers volunteered to participate. One teacher chose not to participate in the project from the beginning. One additional teacher did not submit material to be evaluated at the end of the year.
Patrick County Public Schools	
Blue Ridge Elementary School	All teachers were required to participate.
Hardin Reynolds Memorial School	All teachers were required to participate.
Roanoke City Public Schools	
Patrick Henry High School	All teachers were required to participate.

Table 2 provides a summary of the level of participation in each of the SIG pilot schools, as reported by school principals or the division contact person:

Table 2. Level of Participation: SIG Schools

Colonial Beach Public Schools	
Colonial Beach High School	Teachers were not required to participate, and three out of 18 teachers declined to participate.
Fluvanna County Public Schools	
Central Elementary School	Teachers were not required to participate. Thirty-four participated, and 73 teachers declined participation.
Columbia Elementary School	Teachers were not required to participate. Four participated, and five declined participation.
Cunningham Elementary School	Teachers were not required to participate. Fourteen participated, and six teachers declined participation.
Franklin City Public Schools	
Franklin High School	No teachers were required to participate in the pilot. Four participated, and six teachers declined participation.
Hopewell City Public Schools	
Hopewell High School	All teachers were required to participate.
Northampton County Public Schools	
Kiptopeke Elementary School	Teachers were not required to participate. Three teachers participated.
Northampton High School	The program was not a requirement. Four teachers asked to be part of the pilot.

Petersburg City Public Schools	
A.P. Hill Elementary School	Ten teachers decided to be in the pilot. Of those that participated, one dropped out.
Peabody Middle School	Teachers were not required to participate. Seven teachers participated.
Richmond City Public Schools (Teachers were strongly encouraged to participate in the pilot program in the Richmond City schools. Very few declined to participate. Those who declined had valid reasons for not participating, which included illness, being on leaves, and working as in itinerant teacher in two or more schools.)	
Armstrong High School	All teachers were encouraged, but not required to participate. Seven out of 64 teachers declined to participate.
Boushall Middle School	All teachers were encouraged, but not required to participate. Not all teachers participated. Eleven out of teachers declined to participate, but a variety of content area teachers and elective teachers participated.
Fred Thompson Middle School	All teachers were encouraged, but not required to participate. Four out of 49 teachers declined to participate.
Roanoke City Public Schools	
Fleming High School	All teachers were encouraged, but not required to participate. Some teachers declined to participate.
Lincoln Terrace Elementary School	All teachers were encouraged, but not required to participate, and all of them participated.
Westside Elementary School	All teachers were encouraged, but not required to participate, and all of them participated.

Technical Support Provided to Participating Schools

Staffing and Responsibilities

The College of William and Mary research team was contracted to provide the consultancy and research for this initiative. This team has extensive experience in researching, developing, and supporting the design and application of teacher, leader, and specialist evaluation systems. Additionally, they work extensively on the related issues of teacher effectiveness and consult with educational organizations throughout the world. The following list provides roles that each of the team members engaged in during the pilot.

James H. Stronge, Ph.D. – Project Director: Dr. Stronge is the Heritage Professor of Education, a distinguished professorship, in the Educational Policy, Planning, and Leadership Area at The College of William and Mary, Williamsburg, Virginia. His research interests include policy and practice related to teacher effectiveness, and teacher and administrator evaluation. Dr. Stronge has presented his research at state, national, and international conferences. Additionally, he has worked extensively with state education agencies, local school divisions, and other educational organizations on issues related to teacher effectiveness, and teacher and administrator evaluation. Dr. Stronge was the project director for the Virginia teacher evaluation and performance-pay grant at William and Mary. He served as the primary investigator for the project and provided leadership. His role included, but was not limited to, the following responsibilities:

- serving as a main contact between schools/school divisions and William and Mary;
- providing research on teacher effectiveness and evaluation;
- presenting at conferences, workshops, and professional development trainings;
- organizing and leading planning for design and delivery of division implementation of teacher evaluation; and
- developing training materials associated with implementation of teacher evaluation.

Patricia Popp, Ph.D. – Project Coordinator: Dr. Popp provided organizational guidance, budget management, staffing oversight, materials review, and assistance with delivery of selected professional development activities related to the project. She provided direct support to Richmond City, Franklin City, Fluvanna County, and Hopewell City schools.

Leslie Grant, Ph.D. – Professional Development Coordinator: Dr. Grant's role included the design, implementation, and evaluation of professional development activities associated with the project. Dr. Grant was instrumental in the development of the evaluation instruments. Dr. Grant changed her role in 2011-2012 and provided consulting services related to decisions for student achievement goal setting.

Lauri Leeper, Ph.D. – Research Associate: Dr. Leeper took the lead in developing the student achievement goal-setting workbook and summative evaluation training (January, July, and August 2012 trainings), assisted with research and material development, and provided direct support to all HTS schools and SIG schools in Fluvanna County, Northampton County, and Petersburg City.

Virginia Tonneson, Ph.D. – Research Associate: Dr. Tonneson assisted in the development and review of training materials and provided direct support to Colonial Beach High School.

Kate Wolfe, doctoral candidate – Research Associate: Ms. Wolfe joined the team in April 2012 and assumed direct support for Richmond City and Franklin City schools. She assisted with research and development of training materials.

Xianxuan Xu, Ph.D. – Research Associate: Dr. Xu conducted background research and assisted in the development of research briefs and training materials.

Research associates noted above had the following key responsibilities:
The primary responsibilities of the four research associates included assisting with:

- conducting background research on areas related to the scope of the project;
- examining best practices of existing programs that are consistent with the project deliverables;
- developing research briefs and publications related to the project deliverables;
- developing professional development materials required as part of the professional development activities in the project;
- organizing plans and materials for professional development activities;
- providing support during the implementation of the professional development activities;
- participating in pilot activities for the various target areas of the project;
- designing and selecting instruments for the Teacher Quality Audit;
- collecting data for the various phases of the project;
- providing regular reports to the Virginia Department of Education regarding the progress of the project; and
- other duties as required to fulfill the scope of the project.

Training Materials

Phase I Training Materials

Phase I training materials were designed for use at both the school division and the school level. The training materials were intended to help school divisions and participating schools in aligning their current evaluation systems with the revised *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*. Additionally, the training materials provided practice in implementing a teacher evaluation system that is aligned with the *Guidelines* through simulations and activities. The training materials were organized using the structure in the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*, including the organization of the materials into six parts (Table 3). The training materials included fact sheets and research briefs for teacher evaluation based on current research and best practices. A variety of user-friendly activities also were provided that could be easily used by local school divisions. The manual was organized in a sequential and easy-to-assess format. With this manual, users were able to select the fact sheets, training activities, and other resources based on their local needs.

Table 3. Phase I Training Materials

Part	Description
Part 1 - Introduction	An introduction to the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*
Part 2 – Uniform Performance Standards for Teachers	An overview of the seven Virginia performance standards and the use of performance standards and indicators in the data collection and evaluation rating process
Part 3 – Documenting Teacher Performance	A focus on the required and recommended data sources for teacher evaluation
Part 4 – Connecting Teacher Performance to Academic Progress	Recommendations for including measures of academic progress in a teacher evaluation system
Part 5 – Rating Teacher Performance	Training materials that focus on making summative decisions in teacher evaluation
Part 6 – Improving Teacher Performance	Guidance in the need for and implementation of a performance improvement plan

Each part of the training materials was organized in a similar fashion:

- Explanation of materials – Materials that provide background explanation for the targeted/specific element related to the guidelines
- Activities – Training activities along with directions for use
- Samples – Sample completed forms in some sections, if appropriate
- Simulations – Simulations for implementing the guidelines, in some sections, if appropriate
- Briefs – Briefs that provide background and explanatory material related to various facets of the revised guidelines
- Additional Resources – An annotated listing of additional helpful resources

Phase II Training Materials

Phase II training materials were developed for the 2012 Summer Teacher Evaluation Institutes that were held at nine different locations throughout the Commonwealth during June, July, and August of 2012. The training materials were developed to provide technical assistance and professional development to Virginia's school divisions in the implementation of the Board of Education's recommended model teacher evaluation system. The training materials were intended to provide support to central office supervisory personnel, principals, and teachers in the implementation of the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers* with specific assistance provided in the evaluation of Standard 7 – Student Academic Progress.

Each of the 2012 Summer Teacher Evaluation Institutes was held for two and one-half days, and the training materials were organized using a Day 1, Day 2, and Day 3 format as indicated in the table. Materials were developed for use in a train-the-trainer model. The materials were designed for use by both the trainers and the participants. School division personnel were encouraged to use both the trainer materials and the participant materials as they replicate the training provided at the Institutes.

Making Summative Decisions Guidebook

A *Making Summative Decisions Guidebook* was designed and provided to participating school divisions. This document was developed to provide guidance to school divisions in making summative decisions to include the following topics.

- ***Guidelines for Rating Standard 7: Student Academic Progress*** – This section provided guidance on using multiple data sources related to student academic progress to make a summative rating for Standard 7: Student Academic Progress. Specifically, guidance was provided in the following areas:
  - *Using Student Growth Percentiles in teacher performance evaluation;*
  - *Using student achievement goal setting in teacher performance evaluation; and*
  - *Making a summative rating for Standard 7: Student Academic Progress* to include multiple data sources.

- ***Guidelines for Making Summative Decisions*** – This section provided guidance on making summative decisions using the recommended four-level rubrics for performance standards and arriving at an overall summative rating.

Student Goal Setting Guidebook

A *Student Goal Setting Guidebook* was developed to provide assistance to pilot schools in incorporating student goal setting as a data source for measuring Standard 7: Student Academic Progress. Also, this guidebook helped teachers develop SMART (Specific, Measureable, Appropriate, Realistic, and Time bound) goals.

Additional Training Materials

Additional training materials were developed to support pilot schools:

- Web-based resources were developed and posted on the Virginia Department of Education Web site and a wiki site for local use in training administrators and instructional trainers. The wiki site was a one-stop shop that congregated all the training handouts that were developed.
- Simulation training materials were developed to provide evaluators the opportunity to review and evaluate documents related to a teacher's performance, to annotate observational evidence during a videotaped teacher observation, and to devise a final summative rating. By practicing with simulations, evaluators gained greater understanding of the performance standards and the use of multiple data sources in rating a teacher's performance.
- A library of sample student achievement goal setting forms was compiled. This library was organized by elementary, middle, and high school level. Multiple subjects were included. Most of the goals were developed by the teachers in the pilot and evaluated by

their administrators. Further, they were evaluated by the William and Mary team. All were deemed SMART.

- Handouts were developed for each training session, which included PowerPoint Presentations, activities, and evaluation tools that participants can take and use in their school divisions. They were also disseminated through the William and Mary wiki site and USB flash drives.
- The *Yourtown Teacher Evaluation Handbook* was developed and disseminated which included all the guidance and forms needed for immediate implementation of the new evaluation system.

Statewide Training

The College of William and Mary team provided a series of training to the superintendents, principals, evaluation personnel, and teachers from pilot schools. Formats of the training included three days of summer training, one-day professional development sessions, webinars, face-to-face and online questions and answers sessions. These sessions trained administrative staff, key instructional leaders, and teachers in more effective teacher evaluation procedures and methods through the application of the new Virginia Uniform Performance Standards and teacher evaluation prototype. These training sessions were designed to expand the capacity of administrators to tie student achievement with the formative and summative evaluations of teacher performance. Training included, but was not limited to, the following topical areas:

- Teacher and administrator orientation;
- How to implement the performance evaluation system for more effective evaluation methods and procedures;
- How to use student growth models and student achievement goal setting as one measure in evaluating performance;
- Using performance appraisal rubrics for judging quality of performance; and
- Inter-rater reliability in evaluating teacher performance.

Table 4. Statewide Training Sessions

Date	**Topic**	**Location**
July 27-29, 2011	Initial Evaluation Training for administrators (SIG, HTS, and requested – Alexandria and Newport News School Divisions)	The College of William and Mary
August 2-4, 2011	Initial Evaluation Training for administrators (repeat)	The College of William and Mary
October 11, 2011	One-day follow up training: Virginia Department of Education update, progress in pilot, review and critique of goal setting	The College of William and Mary
November 1, 2011	Webinar for Elementary Teachers	Online

Date	Topic	Location
November 2, 2011	Webinar for Secondary Teachers	Online
November 10, 2011	Student Growth Percentile Webinar	Online
January 26, 2012	Introduction to Summative Decision Making	The College of William and Mary
February 21 and 24, 2012	Virginia's Teacher Evaluation System: Making Summative Decisions; Virginia Association of School Superintendents	Roanoke
May 4, 2012	Webinar: Making Summative Decisions	Virginia Department of Education hosted with Dr. James Stronge presenting
July 19, 2012	Making Summative Decisions	The College of William and Mary
August 6, 2012	Making Summative Decisions (repeat)	The College of William and Mary
September 28, 2012	Virginia's Teacher and Principal Evaluation; [Virginia University Approved Teacher Programs Cohort]	Richmond Virginia Department of Education hosted with Dr. Stronge presenting
October 3, 2012	Virginia's Teacher and Principal Evaluation; [The Virginia Professor of Educational Leadership (VPEL)]	Newport News
October 9, 2012	Virginia Determining Summative Ratings Training	The College of William and Mary
November 1, 2012	Student Growth Percentile Webinar	Online

Local Trainings/Meetings

After the initial training, the William and Mary team continued providing ongoing support throughout the pilot year for pilot schools and school divisions on their implementation of the new teacher evaluation program. The team conducted a number of site visits so that the teachers and evaluators in the pilot schools could receive customized technical support and better understand the components of the evaluation system. During these site visits, the William and Mary team provided applicable information that was consistent with the information provided in the statewide training sessions. The William and Mary team communicated with the pilot schools about the new evaluation systems using individual, department, and faculty-wide meetings. In addition, the William and Mary team sent e-mail updates regularly, distributed hard copies and electronic copies of informational materials such as guidebooks and handbooks, and directed teachers to the Web site resources. The William and Mary team was responsive to the questions posed by the pilot schools through e-mails and phone calls. Through these multiple venues, clear and consistent communication was established.

Local Training Provided to HTS Schools

Targeted technical assistance provided during site visits in HTS schools included the following:

- Observations with principals/assistant principals and, where available, ratings assigned to Virginia's teacher performance standards;
- Feedback provided to principals on the quality of student achievement goals; and
- Selected student achievement goals collected for inclusion in a handbook of exemplars.

Observations

Nine on-site visitations were provided to the identified HTS pilot schools. In each of the school visitations, at least two classroom observations occurred. Each observation was conducted with the school's principal, an assistant principal or, in one case, both the principal and two assistant principals. Twenty classroom observations occurred over the course of six weeks. Each observation lasted a minimum of 40 minutes. In all cases, ratings were generated and calibrated. Inter-rater reliability in the classroom observations was established to a high degree.

Table 5. Site Visits: HTS Schools

Date	William and Mary Staff	Location – school sites	Observers	Number of Observations
February 15, 2012	Leeper	Accomack County – Kegotank Elementary School	Principal	2
February 16, 2012	Leeper	Accomack County – Pungoteague Elementary School	Principal	3
February 27, 2012	Leeper	Caroline County – Madison Elementary School	Principal	2
February 28, 2012	Leeper	Caroline County – Caroline High School	Assistant Principal	2
February 29, 2012	Leeper	Dinwiddie County – Dinwiddie Middle School	Principal and 2 Assistant Principals	2
March 1, 2012	Leeper	Greensville County – Edward W. Wyatt Middle School	Principal	2
March 28, 2012	Leeper	Patrick County – Blue Ridge Elementary School	Principal	3
March 29, 2012	Leeper	Patrick County – Hardin Reynolds Memorial School	Principal	2
March 31, 2012	Leeper	Roanoke City – Patrick Henry High School	Assistant Principal	2

Student Achievement Goal Setting

In addition to multiple observations, at each of the nine schools, at least five student achievement goals were submitted prior to the visit. These were evaluated and specific feedback was given about each of the goals during the visit. Areas of feedback included 1) overall comments, 2) baseline data, 3) goal statement, and 4) possible instructional strategies. These areas were discussed in detail following observations during conferences held between the technical advisor and the principals/assistant principals.

Local Training Provided to SIG Schools

Table 6 provides a summary of the site visits that were conducted in SIG schools.

Table 6. Site Visits: SIG Schools

Date	Topic	Staff	Location – school sites
August 22, 2011	Schools introduction pilot training	Popp and Leeper	Roanoke City – Lincoln Terrace Elementary and Fleming High
August 25, 2011	School introduction pilot training	Leeper	Roanoke City – Westside Elementary
August 25, 2011	Schools introduction pilot training	Popp and Leeper	Fluvanna County – Central Elementary, Columbia Elementary, and Cunningham Elementary
August 26, 2011	Schools introduction pilot training	Leeper	Northampton County – Kiptopeke Elementary, and Northampton High
August 29, 2011	Schools introduction pilot training	Popp	Franklin City – Franklin High
August 29, 2011	Schools introduction pilot training	Tonneson	Colonial Beach – Colonial Beach High
September 1, 201	Schools introduction pilot training	Leeper	Petersburg City – AP Hill Elementary
September 15, 2011	Schools introduction pilot training	Popp and Leeper	Fluvanna County – Central Elementary, Columbia Elementary, and Cunningham Elementary
September 19, 2011	Schools introduction pilot training	Popp	Richmond City – Boushall Middle Petersburg City – Peabody Middle
September 20, 2011	Schools introduction pilot training	Popp	Richmond City – Fred Thompson Middle
September 26, 2011	Schools introduction pilot training	Popp	Richmond City – Armstrong High
October 14, 2011	Schools introduction pilot training	Popp	Hopewell City – Hopewell High
October 20, 2011	School follow up	Popp and Leeper	Fluvanna County – Central Elementary, Columbia Elementary, and Cunningham Elementary
January 3, 2012	Planning meeting for Richmond next steps	Popp	Richmond City – Armstrong High
January 5, 2012	Reviewing goals with schools	Popp	Richmond City – Armstrong High
January 9, 2012	Reviewing goals with schools	Popp	Richmond City – Fred Thompson Middle

Date	Topic	Staff	Location – school sites
January 10, 2012	Reviewing goals with schools	Popp	Richmond City – Boushall Middle
January 31, 2012	Joint observations	Popp and Leeper	Fluvanna County – Central Elementary, Columbia Elementary, and Cunningham Elementary
February 2, 2012	Reviewing goals with school	Popp	Franklin City – Franklin High
February 2, 2012	Reviewing goals with school	Leeper	Northampton County – Northampton High
February 3, 2012	Reviewing goals with school	Leeper	Northampton County – Kiptopeke Elementary
March 29, 2012	Reviewing goals with school	Popp	Hopewell City – Hopewell High
April 23, 2012	Planning with administrators	Popp and Wolfe	Hopewell City – Hopewell High
April 24, 2012	Joint observations	Popp and Wolfe	Hopewell City – Hopewell High
May 3, 2012	Meeting with mathematics teachers	Popp and Wolfe	Hopewell City – Hopewell High
May 3, 2012	Follow up with teachers	Popp and Wolfe	Franklin City – Franklin High
May 4, 2012	Joint observations	Popp and Wolfe	Richmond City – Boushall Middle
May 18, 2012	Joint observations	Popp and Wolfe	Richmond City – Fred Thompson Middle and Armstrong High
June 27, 2012	Initial training of administrators	Popp	Hopewell City (provided for full school division)
September 18, 2012	Training for new teachers and review of SMART process for experienced teachers	Wolfe	Richmond City – Boushall Middle
September 19, 2012	Training for new teachers and review of SMART process for experienced teachers	Wolfe	Richmond City – Armstrong High
September 20, 2012	Training for new teachers and review of SMART process for experienced teachers	Wolfe	Richmond City – Fred Thompson Middle
October 2, 2012	Reviewing goals with all teachers	Wolfe	Richmond City – Boushall Middle

Evaluation of the Pilot Study

Evaluation of the Training

The data of participants' perceptions of the effectiveness of the training they received for the pilot were collected from the training sessions in summer 2011 and summer 2012.

Training: July 27-29, 2011, Three-day Initial Evaluation Training for Administrators, SIG Group

38 respondents

Table 7. Perceptions of July 27-29, 2011 Training: Table of Means

	Mean
Overall, the training:	
Was well organized	3.71
Was relevant	3.79
Had interesting and stimulating activities	3.28
Had effective speakers	3.64
Had effective scheduling and pacing	3.69
The training helped me:	
Become more familiar with the new teacher evaluation standards	3.65
Learn about new resources	3.53
Have an opportunity to network and share ideas	3.16
Develop a plan to begin implementing the new system in my school/division	3.34
The resources provided are:	
Well organized	3.79
Relevant	3.82
Useful in explaining the standards	3.71
Useful in guiding the implementation process	3.68
Adaptable for my school/division	3.66
Logistics:	
Registration	3.84
Travel Accessibility	3.78
Meeting accommodations	3.76

Note: Based on a four-point rating scale: Strongly Agree=4, Agree=3, Disagree=2, Strongly Disagree=1

Table 8. Perceptions of July 27-29, 2011, Training: Table of Percentages

	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**	**Not sure /NA**
Overall, the training was:					
Well organized	71.1%	28.9%	0	0	0
Relevant	78.9%	21.1%	0	0	0
Had interesting and stimulating activities	39.5%	55.3%	2.6%	0	2.6%
Had effective speakers	65.8%	34.2%	0	0	0
Had effective scheduling and pacing	71.1%	26.3%	2.6%	0	0
The training helped me:					
Become more familiar with the new teacher evaluation standards	63.2%	34.2%	0	0	2.6%
Learn about new resources	60.5%	36.8%	0	0	2.6%
Have an opportunity to network and share ideas	47.4%	36.8%	7.9%	0	7.9%
Develop a plan to begin implementing the new system in my school/division	44.7%	50%	2.6%	0	2.6%
The resources provided were:					
Well organized	81.6%	15.8%	2.6%	0	0
Relevant	81.6%	18.4%	0	0	0
Useful in explaining the standards	71.1%	28.9%	0	0	0
Useful in guiding the implementation process	76.3%	21.1%	0	0	2.6%
Adaptable for my school/division	65.8%	34.2%	0	0	0
Logistics:					
Registration	81.6%	15.8%	0	0	2.6%
Travel Accessibility	76.3%	21.1%	0	0	2.6%
Meeting accommodations	73.7%	23.7%	0	0	2.6%

Narrative Evaluation Provided by the Participants Regarding the Training
Note: Representative comments included

What worked well?

- Very knowledgeable presenters.
- Working with our team on the specific plans will play a vital role in effective implementation.
- Time to be reflective with district sites.
- Being guided though the Uniform Performance Standards/goal setting was helpful.
- Great to hear the State Superintendent!
- The notebook was well organized with useful information and will be great resources to help with implementing the plan.
- The presentations were interactive.
- Explanation of 1) resources provided (toolkit/binder); 2) text references; 3) breakdown of goal setting; 4) academic progress was detailed.

- 1) Goal setting – good job walking us through that. 2) The binder is excellent. 3) Electronic access to documents.
- 1) Communication concerning the connection between teacher and principal collaboration was excellent. 2) Resources were detailed and well written.
- Great organization and focus.
- 1) Having time to work with my team on how to implement, what it will look like, etc. 2) Very well organized, efficiently run.
- The presentation was concrete and easy to follow. Great care was given to make the integration of new system seamless and time manageable.
- This seminar provided an excellent balance of input and processing. The resources provided are outstanding and reflect much thought and organization. This was an outstanding seminar. Thanks also for time to work with team.
- Chunk and chew time. I liked given information, then given time to chew on it.
- 1) Clearly defined definition of standards. 2) Ability to adapt observation/evaluation to meet school division needs.
- 1) Samples that were shared/discussed were extremely helpful in understanding the process. 2) Speakers were extremely helpful and "open" to questions, concerns, comments.
- Format – it was good to go back to some documents more than once. We had to "roll up our sleeves" with the material. Excellent training – looking forward to the project.
- Format of the three days was appropriate.

What could be improved? Suggestions for improvement?

- Could have provided access to PowerPoint slides online before or during training, rather than post-training.
- The notebook is not bad; just a bit cumbersome and a bit difficult to flip back and forth. Suggestion: Separate training materials from the rest.
- An extra day in order for the division, possibly as a team, to implement the program and devise a rough draft.
- None – training was effective and relevant.
- Trainer on data lacked interaction and engagement – needed additional activities to keep audience engaged.
- Presenter on Day 2 needed to be more interactive. It was difficult to maintain focus. More activities to engage participants.
- Student progress section (percentage breakdown) was still unclear.
- Communication concerning the connection between teacher evaluation and actual pay for performance could have been improved. Suggestion: All stakeholders should meet and talk to ensure all are on the same "sheet of music."
- Length of sessions.
- The after lunch session on Day 2 was very important and enlightening. I could suggest engaging table team in processing and discussing what is shared throughout that time period as opposed to asking whole group processing questions most of the time. I'm so proud that Virginia is moving in the direction presented by Dr. Jonas.
- More time for chunk/chew.

- 1) Earlier notification that a meeting date is needed for teacher training. 2) More thought into how elective teachers fit into pay for performance.
- 1) Ton of material jam-packed into 3 days. 2) Wish the growth model was in place before beginning the teacher evaluation piece.
- Beyond your control – I would have liked to have this training July 2010 for implementation in September 2011.
- We needed school board level people at the training with us to really move the new system further.
- The training came very late for us – I am nervous as I feel we are already behind.
- An aside: Power strips for outlets for the many computers.
- Our entire team needed to be together for this training – we didn't have all attending – some will be here next week.
- Awareness of timelines.
- Someone from the Virginia Department of Education who can answer specific questions needs to be in attendance every day!!!

What additional training or resources would be helpful?

- Nothing!
- Ongoing training on the performance standards/ongoing training on goal setting/maintaining on SGPs/ how to tie evaluation to SGPs to pay for performance.
- Meetings every few weeks to monitor progress.
- 1) Assessment building; 2) percentage of student progress section.
- It would be helpful to bring in Orange County administrators and teachers to give us actual feedback about the progress.
- Development of assessment and rubrics for non-SOL content areas.
- More on performance assessment beyond math/reading.
- Looking forward to the sessions with the teachers and to the ongoing support.
- Goal setting training for everyone (In-depth training).
- Support talking to teachers and school board office staff.
- Much more help in developing/identifying assessment of student progress.
- Additional training needed in the schools to help sell the program.

Training: August 2-4, 2011, Three-day Initial Evaluation Training for Administrators, Hard-to-Staff Group

32 Respondents

Table 9. Perceptions of August 2-4, 2011, Training: Table of Means

	Mean
Overall, the training:	
Was well organized	3.84
Was relevant	3.91
Had interesting and stimulating activities	3.41
Had effective speakers	3.72
Had effective scheduling and pacing	3.22
The training helped me:	
Become more familiar with the new teacher evaluation standards	3.84
Learn about new resources	3.63
Have an opportunity to network and share ideas	3.50
Develop a plan to begin implementing the new system in my school/division	3.34
The resources provided were:	
Well organized	3.81
Relevant	3.84
Useful in explaining the standards	3.78
Useful in guiding the implementation process	3.75
Adaptable for my school/division	3.50
Logistics:	
Registration	3.50
Travel Accessibility	3.59
Meeting accommodations	3.59

Note: Based on a four-point rating scale: Strongly Agree=4, Agree=3, Disagree=2, Strongly Disagree=1

Table 10. Perceptions of August 2-4, 2011, Training: Table of Percentages

	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**	**Not sure/NA**
Overall, the training:					
Was well organized	84.4%	15.6%	0	0	0
Was relevant	90.6%	9.4%	0	0	0
Had interesting and stimulating activities	50%	46.9%	3.1%	0	0
Had effective speakers	71.9%	28.1%	0	0	0
Had effective scheduling and pacing	53.1%	34.4%	3.1%	0	9.4%
The training helped me:					
Become more familiar with the new teacher evaluation standards	84.4%	15.6%	0	0	0
Learn about new resources	62.5%	37.5%	0	0	0
Have an opportunity to network and share ideas	50%	50%	0	0	0
Develop a plan to begin implementing the new system in my school/division	53.1%	40.6%	0	0	6.3%
The resources provided were:					
Well organized	81.3%	18.8%	0	0	0
Relevant	84.4%	15.6%	0	0	0
Useful in explaining the standards	78.1%	21.9%	0	0	0
Useful in guiding the implementation process	75%	25%	0	0	0
Adaptable for my school/division	68.8%	25%	0	0	6.3%
Logistics:					
Registration	71.9%	18.8%	3.1%	0	6.3%
Travel Accessibility	65.6%	28.1%	6.3%	0	0
Meeting accommodations	78.1%	15.6%	0	0	6.3%

Narrative Evaluation Provided by the Participants Regarding the Training

What worked well?

- Location, facility, resources.
- 1) Organization of the overall evaluation plan and the way it was presented worked well. 2) Opportunities for audience to ask questions worked well.
- 1) Breakout activities. 2) Small group discussion. 3) Share-outs.
- The transitions and the simulated learning.
- The information was well prepared and explained. It was made available so that I can use it with my staff.
- Affirmation on the direction we were already headed.
- Having sufficient research completed prior to the conference.
- 1) Knowledgeable presenters. 2) Connection of content to what is expected in schools. 2) Support offered/provided.

- The overall process and organization was excellent. The resources provided were amazing and extremely relevant. Love the “new School of Education.” Thank you for having the Dean stop by and share her insight not only from her position here but also as a Virginia Board of Education member.
- 1) Presentations. 2) Explanation of the new “evaluation”: standards, expectations, and consequences.

What could be improved? Suggestions for improvement?

- Three days could be reduced to two.
- Perhaps allow for more breaks to get up and move around. Hard time focusing when sit for long periods of time (2+ hours).
- Ensure that all presenters engage all participants (especially after lunch).
- The section on student growth measures should be more interactive.
- Maybe in later training, break out by elementary, middle, high to facilitate more networking and level specific issues.
- The location is far from southwest Virginia.
- Receiving more information prior to the conference in order to wrap our minds around what will be discussed.
- Shorten the time between breaks.
- Activities that would allow us more time to network with other divisions and share ideas.
- Timeframe for implementation.
- The amount of information was overwhelming especially for those who are expected to implement this process in a few weeks. More preparation time would be greatly appreciated – but not possible at this point.

What additional training or resources would be helpful?

- Evaluation of goal setting process. Support in defining and recognizing goals across subject areas. We did some of this and we are supposed to do more in October. Thanks for support!
- Concerns about implementation with fidelity if certain measures are still not in place.
- More information on identifying the “exemplary” teacher. Training/workshops for teacher leaders.
- Quicker turn around in follow-up training.
- Clarification on summative rating.
- 1) Training and workshops for “teacher-leaders.” 2) I am excited about the vision that Virginia Department of Education has for the evaluation system and the support of the experts from William and Mary. This is long overdue. Just a little anxiety about some of the “what-ifs.”
- More training on goal setting.
- Additional training on components of the new evaluation plan.

Training: July 19, and August 6, 2012, One-day Summative Decisions Training

72 Respondents (including both School Improvement and Hard-to-Staff groups)

Table 11. Perceptions of July 19 and August 6, 2012, Training: Table of Means

	Mean
Overall, the training:	
Was well organized	3.63
Was relevant	3.61
Had interesting and stimulating activities	3.39
Had effective speakers	3.55
Had effective scheduling and pacing	3.52
The training helped me:	
Become more familiar with the new teacher evaluation standards	3.60
Learn about new resources	3.41
Have an opportunity to network and share ideas	3.28
Develop a plan to begin implementing the new system in my school/division	3.36
The resources provided were:	
Well organized	3.64
Relevant	3.69
Useful in explaining the standards	3.61
Useful in guiding the implementation process	3.60
Adaptable for my school/division	3.50
Logistics:	
Registration	3.63
Travel Accessibility	3.59
Meeting accommodations	3.64

Note: Based on a four-point rating scale: Strongly Agree=4, Agree=3, Disagree=2, Strongly Disagree=1

Table 12. Perceptions of July 19 and August 6, 2012, Training: Table of Percentages

	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**	**Not sure /NA**
Overall, the training:					
Was well organized	62.5%	36%	0	0	1.4%
Was relevant	61.1%	38.8%	0	0	0
Had interesting and stimulating activities	40.2%	48.6%	4.2%	0	6.9%
Had effective speakers	54.2%	44.4%	0	0	1.4%
Had effective scheduling and pacing	50.0%	45.8%	0	0	4.2%
The training helped me:					
Become more familiar with the new teacher evaluation standards	58.3%	38.9%	0	0	2.8%
Learn about new resources	36.1%	41.7%	2.8%	0	19.4%
Have an opportunity to network and share ideas	34.7%	44.4%	9.7%	0	11.1%
Develop a plan to begin implementing the new system in my school/division	33.3%	37.5%	5.6%	0	23.6%
The resources provided were:					
Well organized	63.9%	36.1%	0	0	0
Relevant	68.1%	30.6%	0	0	1.4%
Useful in explaining the standards	59.7%	37.5%	0	0	2.8%
Useful in guiding the implementation process	55.6%	37.5%	0	0	6.9%
Adaptable for my school/division	62.5%	37.5%	0	0	0
Logistics:					
Registration	62.5%	33.3%	1.4%	0	2.8%
Travel Accessibility	59.7%	36.1%	1.4%	0	2.8%
Meeting accommodations	62.5%	34.7%	0	0	2.8%

Narrative Evaluation Provided by the Participants Regarding the Training

What worked well?

- The training on the decision making process for Standard 7 was very good. Presenters did an excellent job. Kate Wolfe was very good at answering questions.
- The materials were easily accessible and will be greatly utilized.
- Best explanation to date.
- Pacing and group activities.
- Shorter segments.
- The simulations were very helpful.
- The training was very organized, informative, and helpful in further understanding the process.
- Excellent manual to assist with summative decision making in the field.

What could be improved?

- Wording on some forms.
- Summative decisions.
- Continue to clarify with more examples for a deeper understanding of proficient versus exemplary.
- Pacing.
- Speakers speak more slowly and provide more explicit directions.
- Lunch.
- Organization of resources into one packet that is easier to use.
- Shorter videos. The practice activities were too long.
- A more central location rather than travelling across the state.
- Structured facilitated sharing time with other divisions.
- More simulations.
- Movement and interactive activities.
- Provide more opportunities to review our own materials with colleagues.
- Would like a section in the book with extra "forms" to use during real evaluations!

Suggestions for improvement?

- More debate on the ratings so that people could hear the decision-making process more.
- More practice.
- More interactions for participants.
- Divide by high school or secondary and elementary.
- Workshop should be limited to ½ day.
- Cut down the 27 minute video.
- Use our own information – some time spent on talking with our division people.

Summary of Evaluation of Training

Evaluators need training in order to develop the knowledge and skills they need to implement a new evaluation system. The participants were surveyed about their experience with major statewide training sessions that were provided. The survey data indicated that the participants consistently had positive perceptions that the training was well organized and relevant, and training had interesting and stimulating activities. They also agreed that the training had effective speakers and effective scheduling and pacing. They perceived that the training helped them become familiar with the new teacher evaluation system. They believed they learned about new resources and had opportunities to network and share ideas. The resources provided were perceived as well organized, relevant, useful in guiding the implementation process, and adaptable for local school/division use. The participants perceived the speakers were helpful and open to questions, concerns, and comments.

The survey data also indicate there is a need for additional training. Some participants expressed they needed more in-depth and ongoing training on the performance standards, Student Growth Percentiles (SGPs), and goal setting. Some said they would like to have meetings every few weeks to monitor progress. Some would like to receive external evaluations of their goal setting process (which was provided later through site visits and goals reviews). Some also shared that there is a need to improve logistics of the training, for instance, separating the training session by secondary and elementary schools, and sending the information prior to the training so that the attendees have opportunities to familiarize themselves with what will be discussed. Some shared concerns about the implementation with fidelity and needed follow-ups.

Inter-rater Reliability

To ensure the trustworthiness of the teacher evaluation system model, it is important to check the inter-rater reliability and validity of trained evaluators. In order to check their inter-rater reliability, the evaluators participated in simulation activities in the trainings where they evaluated teachers in video-based simulations and completed the *Teacher Interim Performance Report* based on the data of teacher performance that was provided. The tested Virginia teacher evaluation system has four levels of performance: Exemplary, Proficient, Developing/Needs Improvement, and Unacceptable. Performance rubrics are used to guide the summative rating of each of the seven standards. The performance rubric is a behavioral summary scale that describes performance levels for each of the seven teacher performance standards. They are provided to increase reliability among evaluators and to help teachers focus on ways to enhance their teaching practice. The highlighted cells in the Tables 13-18 indicate the percentage of evaluators who gave the identical ratings as the intended ratings established by the William and Mary team.

January 26, 2012, Training

This training session provided a simulation activity. In this activity, the participants were asked to work individually to: 1) watch the teacher simulation video; 2) review the evidence from the additional observation; 3) review the documentation as it related to their assigned standards; and 4) use the performance rubrics and the Training Rating Form to rate their four assigned standards.

Table 13. Rating by Individual Participants for a Teacher in a Simulation Activity

Standard 1	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	36 (82%)	8 (18%)	0	0
Standard 2	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	13 (30%)	30 (68%)	1 (2%)	0
Standard 3	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	4 (9%)	36 (82%)	4 (9%)	0
Standard 4	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	4 (9%)	36 (82%)	4 (9%)	0
Standard 5	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	35(78%)	10(22%)	0	0
Standard 6	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	40 (93%)	3 (7%)	0	0
Standard 7	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	48 (92%)	4 (8%)	0

(Highlighted ratings are the ratings recommended by the William and Mary team.)

February 21, 2012, Training

The participants were asked to watch the teacher simulation video and discuss all of the evidence with their partners and decide upon a summative rating for each standard and post them on the charts hanging on the wall.

Table 14. Rating by Groups for a Teacher in a Simulation Activity

Standard 1	**Exemplary**	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	5 (45%)	6 (55%)	0	0
Standard 2	Exemplary	**Proficient**	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	10 (100%)	0	0
Standard 3	Exemplary	**Proficient**	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	9 (100%)	0	0
Standard 4	Exemplary	**Proficient**	Developing/Needs Improvement	Unacceptable
Number (Percentage)	1 (12.5%)	7 (87.5%)	0	0
Standard 5	Exemplary	**Proficient**	Developing/Needs Improvement	Unacceptable
Number (Percentage)	3 (27%)	8 (73%)	0	0
Standard 6	**Exemplary**	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	8 (100%)	0	0	0
Standard 7	Exemplary	**Proficient**	Developing/Needs Improvement	Unacceptable
Number (Percentage)	1 (5%)	18 (95%)	0	0

(Highlighted ratings are the ratings recommended by the William and Mary team.)

February 24, 2012, Training

The participants were asked to watch the teacher simulation video and discuss all of the evidence with their partners and decide upon a summative rating for each standard and post them on the charts hanging on the wall.

Table 15. Rating by Individual Participants for a Teacher in a Simulation Activity

Standard 1	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	3 (23%)	10 (77%)	0	0
Standard 2	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	3 (25%)	9 (75%)	0	0
Standard 3	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	2 (14%)	12 (86%)	0	0
Standard 4	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	2 (15%)	9 (69%)	2 (15%)	0
Standard 5	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	4 (29%)	9 (64%)	1 (7%)	0
Standard 6	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	8 (50%)	8 (50%)	0	0
Standard 7	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	2 (10%)	17 (85%)	1 (5%)	0

(Highlighted ratings are the ratings recommended by the William and Mary team.)

July 19, 2012, Training

The participants were asked to work individually to: 1) watch the teacher simulation video; 2) review the evidence from the additional observation; 3) review the documentation as it related to their assigned standards; and 4) use the performance rubric and the Training Rating Form to rate their four assigned standards.

Table 16. Rating by Individual Participants for a Teacher in a Simulation Activity

Standard 1	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	10 (59%)	7 (41%)	0
Standard 2	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	5 (29%)	8 (47%)	4 (24%)	0
Standard 3	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	1 (5%)	17 (85%)	2 (10%)	0
Standard 4	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	16 (80%)	4 (20%)	0
Standard 5	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	13 (72%)	5 (28%)	0	0
Standard 6	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	14 (70%)	6 (30%)	0	0
Standard 7	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	9 (33%)	18 (67%)	0	0

(Highlighted ratings are the ratings recommended by the William and Mary team.)

July 19, 2012, Training

The participants were asked to discuss all of the evidence with their partners and decide upon a summative rating for each standard and post them on the charts hanging on the wall.

Table 17. Rating by Groups for a Teacher in a Simulation Activity

Standard 1	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	6 (75%)	2 (25%)	0	0
Standard 2	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	3 (37.5%)	5 (62.5%)	0	0
Standard 3	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	5 (83%)	1 (17%)	0
Standard 4	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	4 (67%)	2 (33%)	0
Standard 5	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	5 (62.5%)	3 (37.5%)	0	0
Standard 6	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	7 (87.5%)	1 (12.5%)	0	0
Standard 7	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	12 (86%)	2 (14%)	0

(Highlighted ratings are the ratings recommended by the William and Mary team.)

August 6, 2012, Training

The participants were asked to work individually to: 1) watch the video; 2) review the evidence from the additional observation; 3) review the documentation as it related to their assigned standards; and 4) use the Virginia Training Rating Form to rate their four assigned standards.

Table 18. Rating by Individual Participants for a Teacher in a Simulation Activity

Standard 1	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	8 (42%)	11 (58%)	0	0
Standard 2	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	6 (33%)	11 (61%)	1 (6%)	0
Standard 3	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	0	13 (59%)	9 (41%)	0
Standard 4	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	8 (42%)	11 (58%)	0	0
Standard 5	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	7 (41%)	9 (53%)	1 (6%)	0
Standard 6	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	16 (84%)	3 (16%)	0	0
Standard 7	Exemplary	Proficient	Developing/Needs Improvement	Unacceptable
Number (Percentage)	5 (15%)	28 (82%)	1 (3%)	

(Highlighted ratings are the ratings recommended by the William and Mary team.)

Summary of Inter-Rater Reliability Assessment

These tables (13-18) indicate that inter-rater reliability of the evaluators' ratings was generally high across all of the seven performance standards of the evaluation system. Typically, the ratings awarded by about 60 percent to 100 percent of the individuals and groups had an exact match with the William and Mary team, but in a few cases, the percentage of exact matches was below 50 percent. In these cases, the participants typically rated one level off, either above or below the rating of William and Mary team. There was no case where the participants would score two levels off the target. The standard with the highest percentage of exact matches was Standard 3 – Instructional Delivery, and the standard with the lowest was Standard 5 – Assessment of and for Learning.

Overall, the participants typically awarded the same ratings to the same lessons and teachers as established by the William and Mary team in their calibration of the simulations. This high level of inter-rater reliability indicates that the training was able to build the evaluators' understanding of the components of the new teacher evaluation system, including the Standard 7 – the measure of student academic progress, and the evaluators were able to use the rubrics to guide their evaluation of a teacher's performance on all seven standards. The inter-rater reliability shows that evaluators were able to make summative rating using the new evaluation system with fidelity in the simulation activities. These findings suggest that the evaluators can successfully implement the summative decision process in their schools.

Analysis of the Final Pilot-Year Evaluation Data

Data Collection

Data requests were sent by e-mail to the principals and school division contacts of the nine Hard-to-Staff (HTS) and 16 School Improvement Grants (SIG) schools. This data request asked that the teacher-level data spreadsheets that had been used in the collection of the ratings be uploaded to a secure dropbox location. After six follow-up requests, spreadsheets for all schools were received.

Data Conditioning

The spreadsheet data from each of the schools were merged into a common dataset. During this process it became apparent that there were inconsistencies in the raw data. Some of the data were in the form of letters or words instead of the requested numerical values. Some of the data values were out-of-bounds, for example, zero values were entered for ratings. Standards data for one school division were the averages of sub-ratings rather than ratings on the 1 to 4 scale. Other data were not provided. In some cases this was a single rating value but in other instances this was all ratings.

Where data were easily translated, the corrected values were replaced in the dataset. For example, alpha values were replaced with their appropriate numeric values. There were several attempts to get clarifications or corrections from schools resulting in additional requests for updated data. These efforts did result in reducing the amount of missing information but did not alleviate all of the data issues. Follow-up efforts also revealed that some of the missing information was attributable to non-teaching personnel being included in the spreadsheet. The data for those individuals were removed from the dataset. Data that were not able to be translated after these attempts were set to missing.

The final summary ratings were to be based on the sum of the ratings for each of the standards with Standard 7 – Student Academic Progress being weighted 40 percent. Since the data calculations reported in the schools' spreadsheets were not consistent with the reporting requirements (for instance, some schools reported Standard 7 – Student Academic Progress on a scale of 1 to 4, while some schools reported the weighted scores), the summary rating values were calculated from the standards ratings based on the instructions set forth in the *Guidelines*. This process did not alter the distribution and the variability of the distribution of teachers' evaluation results within the school but did enable the examination of rating across schools.

Results

Data were reported for 782 teachers representing 25 schools and 13 school divisions. The teachers were from elementary (221, 28.3%), middle (142, 18.2%) and high (419, 53.6%) schools. There were 225 (28.8%) teachers who were judged eligible for performance pay. Additionally, the teachers came from schools that were either SIG or HTS. There were 442 teachers from SIG schools of which 100 (22.6%) were deemed eligible for performance pay.

There were 340 teachers from HTS schools of which 125 (36.8%) were deemed eligible for performance pay.

In the following sections, the results for the HTS and SIG schools will be presented separately. Only teachers with ratings on all standards will be considered in these analyses. Of the 782 teachers reported by the schools, 11 (1.4%) had missing data on at least one standard. There were nine (2.0%) teachers from SIG schools and two (0.6%) teachers from HTS schools with missing standard data. For these individuals, full data were not reported to The College of William and Mary team, therefore, they were excluded from this study. The final sample with complete standards ratings was 771 (98.6%). There were 433 (98.0%) HTS teachers and 338 (99.4%) SIG teachers.

HTS School Results

Standards 1 to 7 Descriptive Outcomes

Tables 19 to 25 and Figures 1 to 7 show the results of the ratings for Standards 1 to 7 for the teachers from HTS schools. The tables and figures indicate that Proficient was the most used category for each standard. Between 59 and 75 percent of the ratings for each standard were in this category. Exemplary was the next most often used category with 18 to 37 percent of the teachers being rated in this category per standard. Developing/Needs Improvement and Unacceptable were not often used for many of the standards with the cumulative percentage rated in those two ranges being generally three to six percent. Among the standards, Standard 7 – Student Academic Progress had the most variability of ratings. It is the only standard where considerable numbers were rated Developing/Needs Improvement or Unacceptable (16 percent).

Table 19: Rating Distribution for Standard 1 – Professional Knowledge

Rating	Frequency	Percent
Unacceptable	1	.2
Developing/Needs Improvement	10	2.3
Proficient	261	60.3
Exemplary	161	37.2
Total	433	100.0

Table 20: Rating Distribution for Standard 2 – Instructional Planning

Rating	Frequency	Percent
Unacceptable	1	.2
Developing/Needs Improvement	17	3.9
Proficient	307	70.9
Exemplary	108	24.9
Total	433	100.0

Table 21: Rating Distribution for Standard 3 – Instructional Delivery

Rating	Frequency	Percent
Unacceptable	0	0
Developing/Needs Improvement	29	6.7
Proficient	298	68.8
Exemplary	106	24.5
Total	433	100.0

Table 22: Rating Distribution for Standard 4 – Assessment of and for Student Learning

Rating	Frequency	Percent
Unacceptable	1	.2
Developing/Needs Improvement	22	5.1
Proficient	326	75.3
Exemplary	84	19.4
Total	433	100.0

Table 23: Rating Distribution for Standard 5 – Learning Environment

Rating	Frequency	Percent
Unacceptable	1	.2
Developing/Needs Improvement	18	4.2
Proficient	257	59.4
Exemplary	158	36.4
Total	433	100.0

Table 24: Rating Distribution for Standard 6 – Professionalism

Rating	Frequency	Percent
Unacceptable	2	.5
Developing/Needs Improvement	17	3.9
Proficient	263	60.7
Exemplary	151	34.9
Total	433	100.0

Table 25: Rating Distribution for Standard 7 – Student Academic Progress

Rating	Frequency	Percent
Unacceptable	23	5.3
Developing/Needs Improvement	48	11.1
Proficient	282	65.1
Exemplary	80	18.5
Total	433	100.0

Figure 1: Rating Distribution for Standard 1 – Professional Knowledge

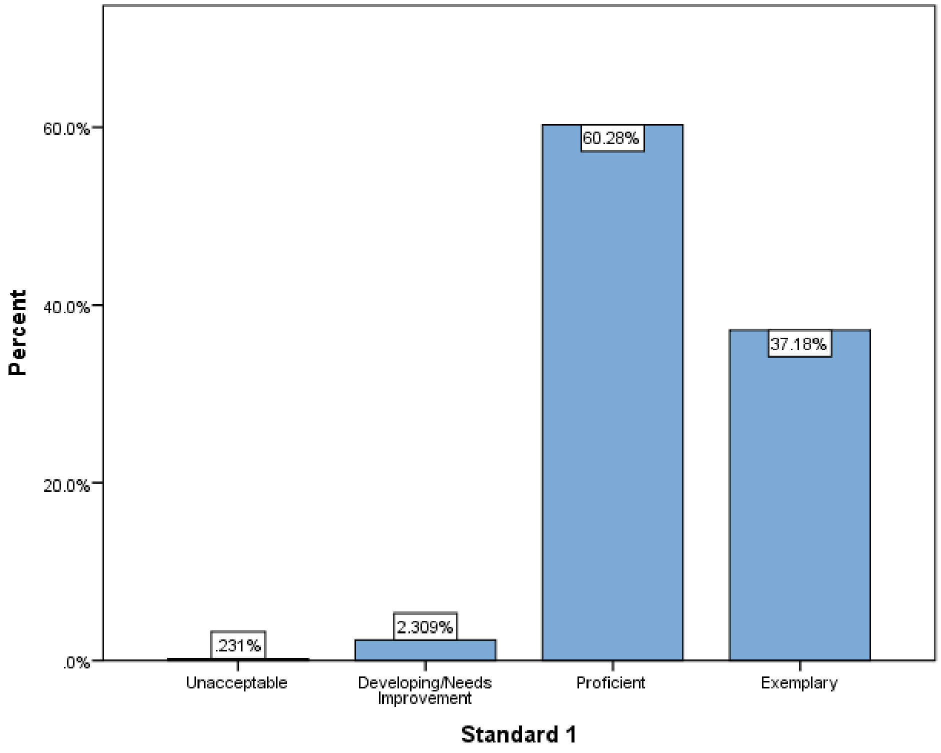


Figure 2: Rating Distribution for Standard 2 – Instructional Planning

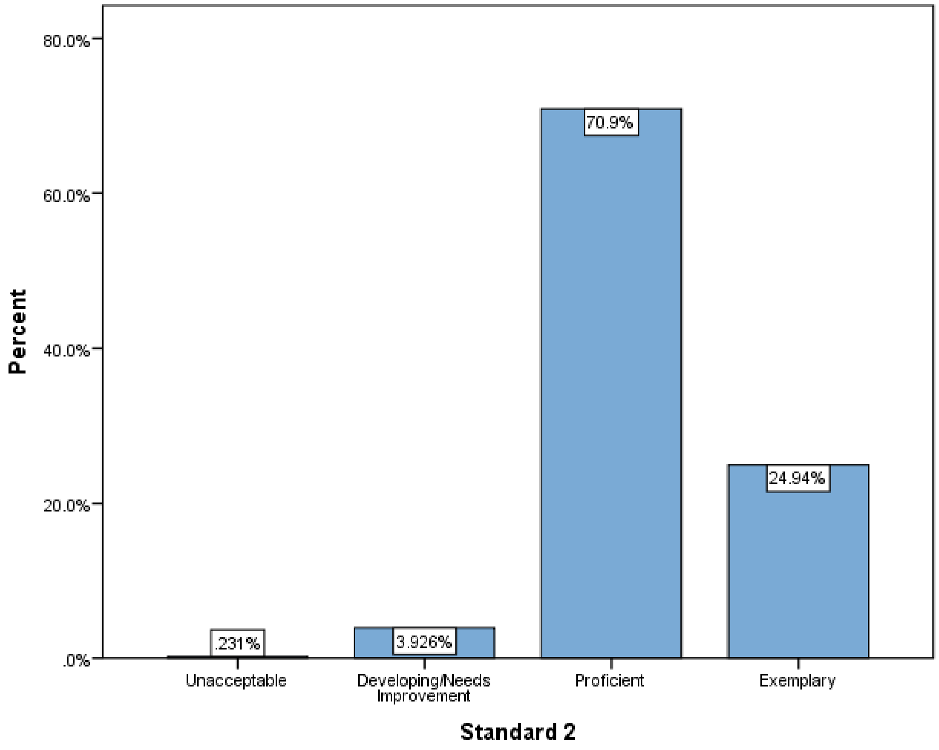


Figure 3: Rating Distribution for Standard 3 – Instructional Delivery

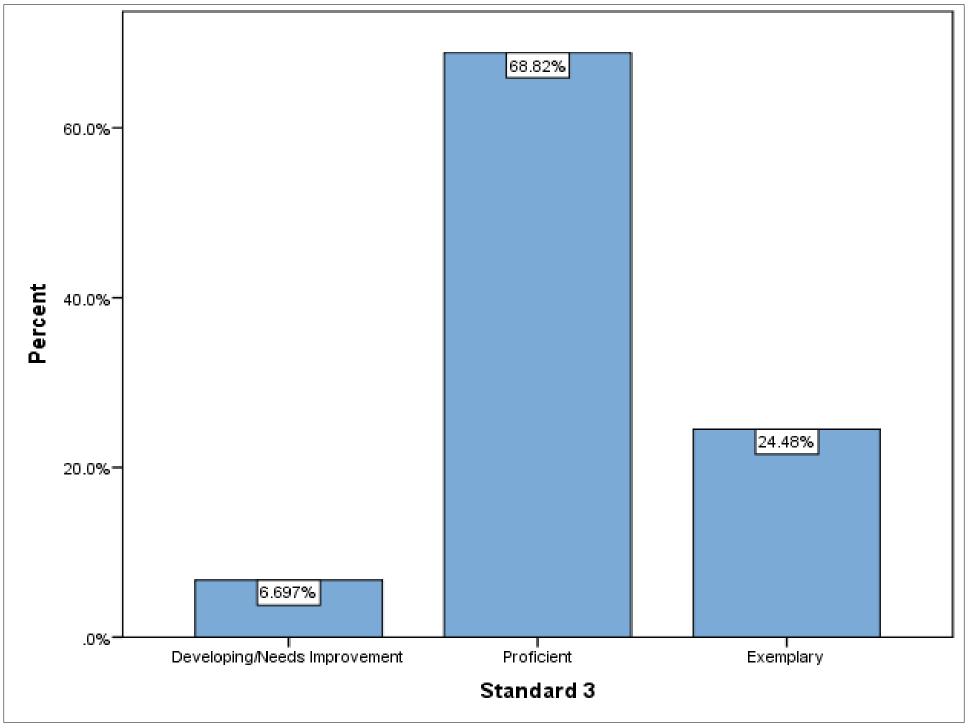


Figure 4: Rating Distribution for Standard 4 – Assessment of and for Student Learning

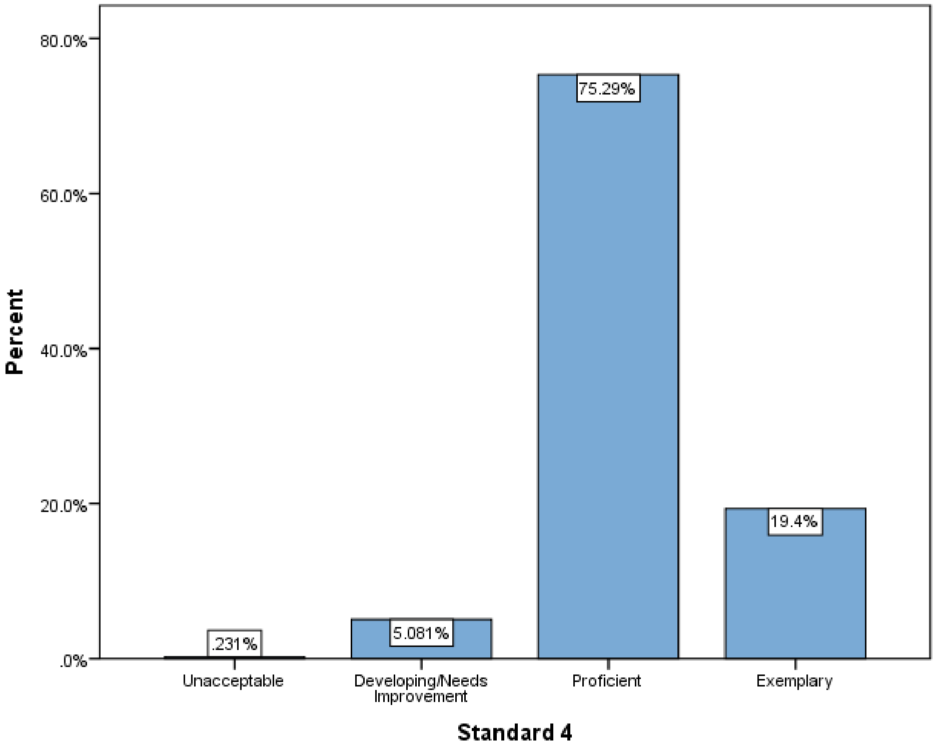


Figure 5: Rating Distribution for Standard 5 – Learning Environment

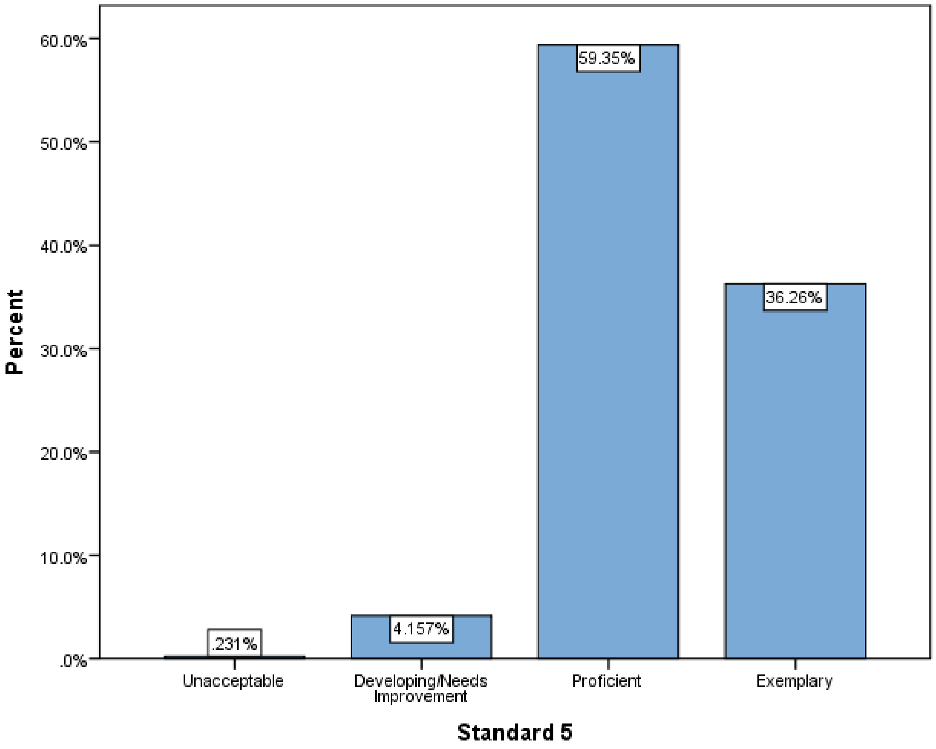


Figure 6: Rating Distribution for Standard 6 – Professionalism

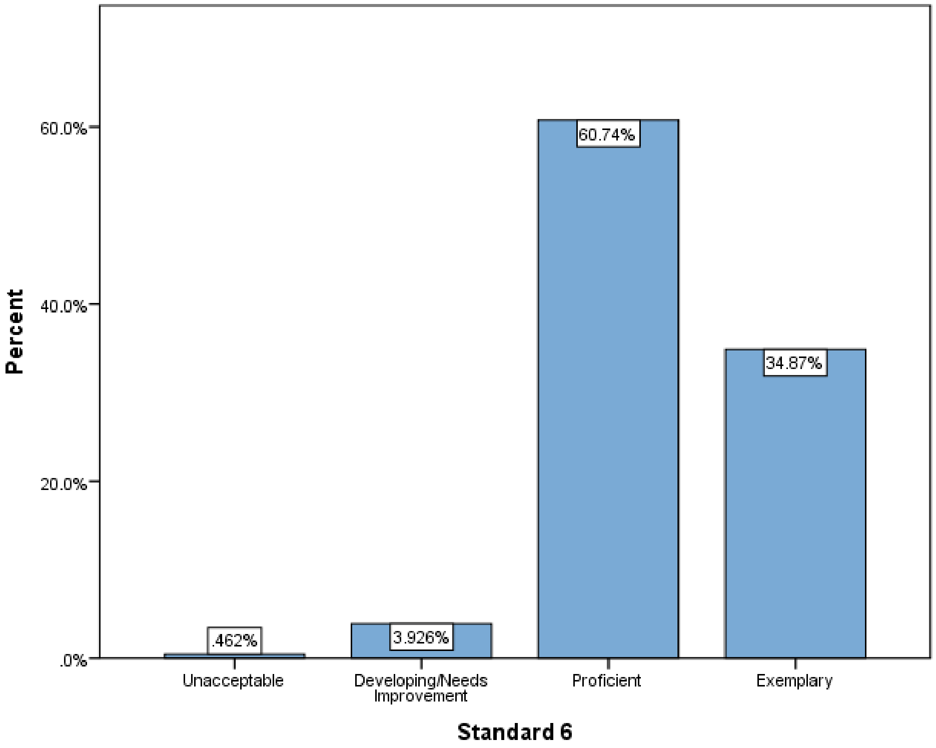
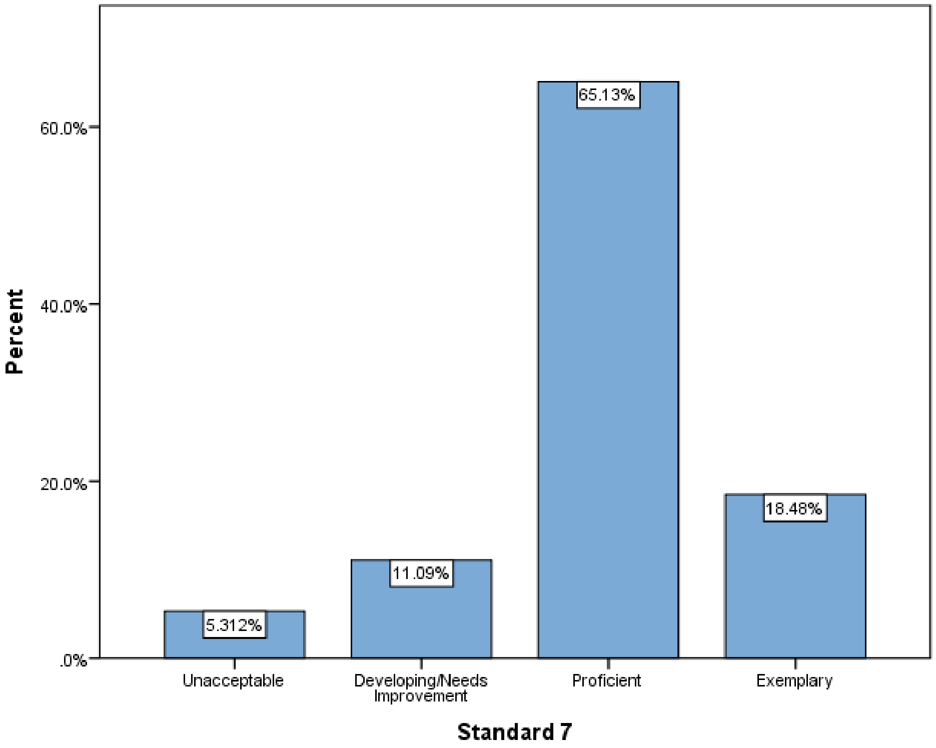


Figure 7: Rating Distribution for Standard 7 – Student Academic Progress



Evaluation Questions

Question 1: What are the relationships among the ratings on the six teacher process standards (Standards 1-6)?

Table 26 provides the correlations among the six process standards (Standard 1 to Standard 6) for the HTS school sample. As can be seen in the table, most of the correlations among the standards are in the moderate range indicating that the ratings of the process standards have considerable overlap. The ratings of Standard 3 – Instructional Delivery and Standard 5 – Learning Environment have the most in common (r = .582); while the ratings of Standard 5 – Learning Environment and Standard 6 – Professionalism are the most unique (r = .351). The correlations in the .4 to .6 range indicate that the standards are rated with some common basis. It may be that an overall impression of the teacher's competence underlies all of the ratings and accounts for the correlation among these standards.

Table 26: Correlations Among the Six Process Standards

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and for Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 1: Professional Knowledge	1					
Standard 2: Instructional Planning	.473**	1				
Standard 3: Instructional Delivery	.518**	.580**	1			
Standard 4: Assessment of and for Student Learning	.565**	.544**	.552**	1		
Standard 5: Learning Environment	.516**	.513**	.582**	.443**	1	
Standard 6: Professionalism	.478**	.465**	.402**	.449**	.351**	1

Note: ** correlation is significant at the .01 level. N = 433.

Question 2: To what degree do ratings of teachers' six process standards predict student academic growth as measured by Standard 7 – Student Academic Progress?

Table 27 displays the correlations between the six process standards and Standard 7 – Student Academic Progress, which is the rating of student academic progress. The correlations in Table 27 are considerably lower than those of Table 26, indicating less overlap between the process standards and student academic progress. All of the correlations are significant and in the moderate range, indicating that there is commonality between all of the process standards and the rating of academic growth. Standard 6 – Professionalism has the weakest relationship to Standard 7 – Student Academic Progress.

Table 27: Correlations Between the Six Process Standards and the Achievement Standard

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and for Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 7: Student Academic Progress	.339**	.339**	.365**	.435**	.356**	.313**

Note: ** correlation is significant at the .01 level. N = 433.

The relationship between the process standards and the academic progress measure was expected to be multivariate. That is, it was expected that a combination of process standard ratings would be required to predict the academic outcome rating. To test this, the six process standard ratings were used as predictors in a stepwise multiple regression with Standard 7, the student academic progress rating, as the target. Table 28 presents the results of the regression analysis. The analysis indicated that process standard ratings for Standards 4, 5, and 6 were the only significant predictors of the student academic progress rating, Standard 7. The overall model multiple R was .482 with an R-square of .232. This indicates a modest ability to predict the Student Academic Progress rating from the process standard ratings. The selection of multiple process standard ratings for the model confirms that the student academic progress measure is multivariate in nature. The modest predictive power of the model indicates that other factors beyond those represented in the six process standards are influential in the student academic progress ratings.

Table 28. Stepwise Regression Results.

Summary of Steps

Model	R	R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
a	.435	.189	2.57	.189	100.37	1	431	.000
b	.471	.222	2.52	.033	18.37	1	430	.000
c	.482	.232	2.50	.010	5.42	1	429	.020

a. Predictors: (Constant), Standard 4
b. Predictors: (Constant), Standard 4, Standard 5
c. Predictors: (Constant), Standard 4, Standard 5, Standard 6
d. Dependent Variable: Standard 7

Variable	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	1.321	.949		1.39	.164
Standard 4	1.784	.296	.304	6.02	.000
Standard 5	.926	.245	.182	3.78	.000
Standard 6	.569	.244	.112	2.32	.020

<u>Question 3</u>: Are the results for teachers with Student Growth Percentiles (SGPs) similar to the results for teachers who have student achievement goal setting as a major measure of student progress?

The Standard 7 – Student Academic Progress ratings of teachers with and without SGPs were compared to determine whether the availability of SGPs had an impact on the Standard 7 ratings. Table 29 displays the mean Standard 7 ratings (weighted) for the two groups. A t-test (***t*** (df = 431) = .174, p = .862) indicated that there was no difference in the mean ratings of the two groups. This finding suggests that there was no differential influence of SGP availability on the Standard 7 ratings. This finding can be considered encouraging in that it suggests evaluators were able to apply both Student Achievement Goal Setting data and Student Growth Percentile data in a manner that yielded comparable ratings for Standard 7.

Table 29. Standard 7 Descriptive Statistics for Teachers With and Without SGPs

SGPs Used	N	Mean	Standard Deviation	Standard Error Mean
No	394	11.88	2.79	.14
Yes	39	11.79	3.43	.55

Although the evaluation question did not directly ask about differences in the process standard ratings, they were tested to see whether differences existed between the SGP and non-SGP groups. Table 30 shows the descriptive statistics and the t-test values for the six process standards by SGPs group. There were significant differences on the first five standards. This indicates that teachers who have SGPs data available were perceived by evaluators as performing better on Standards 1 through 5.

Table 30. Standard 1 through 6 Descriptive Statistics for Teachers With and Without Student Growth Percentiles (SGPs)

Standard	SGPs Used?	N	Mean	Std. Deviation	Std. Error Mean	t value
Standard 1	No	394	3.32	.52	.03	-3.36**
	Yes	39	3.62	.59	.09	
Standard 2	No	394	3.18	.49	.02	-3.35**
	Yes	39	3.46	.60	.10	
Standard 3	No	394	3.15	.51	.03	-3.22**
	Yes	39	3.44	.64	.10	
Standard 4	No	394	3.12	.47	.02	-3.00**
	Yes	39	3.36	.58	.09	
Standard 5	No	394	3.29	.56	.03	-2.61**
	Yes	39	3.54	.55	.09	
Standard 6	No	394	3.30	.55	.03	-0.38
	Yes	39	3.33	.70	.11	

** Groups are significantly different at $p < .01$ level.

Note: The ratings of the standards have been treated here as interval type indicators. The ratings of the standards could also be considered as ordinal indicators. The statistical treatment of such indicators would be different. Analyses were also conducted treating the standard ratings as ordinal data with the results indicating the same conclusions as when the standard ratings were treated as interval data.

Question 4: Do school principals sufficiently discriminate in the application of the teacher evaluation system based on measures of effectiveness?

Consideration of this question requires an examination of the summary ratings. To arrive at a summary rating, raters were instructed to combine the six process standards ratings with the Student Academic Progress Standard (Standard 7). Standard 7 was to be weighted 40 percent and each of the other six standards was to be weighted as ten percent. Examination of the reported data indicated that not all summary ratings conformed to this structure. For instance, some summary ratings were on a scale of 1 to 4, while some were in the form of text (e.g., “Exemplary,” “Proficient,” “Developing/Needs Improvement,” or “Unacceptable”). Therefore, the summary ratings were recalculated based on the approved formula provided in the *Guidelines* to ensure the data were in a consistent format while not changing the principals’ actual ratings. The formula used was: Summary Rating = (Standard 1 Rating × 1) + (Standard 2 Rating × 1) + (Standard 3 Rating × 1) + (Standard 4 Rating × 1) + (Standard 5 Rating × 1) + (Standard 6 Rating × 1) + (Standard 7 Rating × 4). This process did not alter the distribution and the variability of the distribution of teachers' evaluation results within the school but did enable the examination of ratings across schools.

Table 31 shows the descriptive statistics for the summary ratings. Summary ratings could vary between 16 and 40. The table shows that the mean value was toward the higher end of the rating scale with a standard deviation of approximately 5 points. No teachers were rated below 16 but some were rated at the top of the scale. Figure 8 is a graphical representation of the summary ratings. This figure confirms that summary ratings cluster on the higher side of the modal point 30. The figure also confirms that there is variability among the summary ratings but that the lower end of the possible distribution is not utilized as often as the top end.

Summary ratings can be translated into the descriptive categories of the original standards. Values 10 through 19 indicate "Unacceptable"; 20 through 25 indicate "Developing/Needs Improvement"; 26 through 34 indicate "Proficient"; 35 through 40 indicate "Exemplary". Table 32 shows the distribution of summary ratings by category. Based on these percentages, approximately 27 percent of the teachers were rated in the "Exemplary" range and approximately nine percent were rated "Developing/Needs Improvement," or "Unacceptable". Again, this table indicates that discrimination is occurring but mostly at the upper range of the rating scale.

Table 31. Descriptive Statistics for the Summary Rating.

	Range	Minimum	Maximum	Mean	Std. Deviation
Summary Rating	24.00	16.00	40.00	32. 17	5.30

Note: The possible range of scores is 10-40; the actual allocated range was 16-40.

Figure 8. Summary Ratings

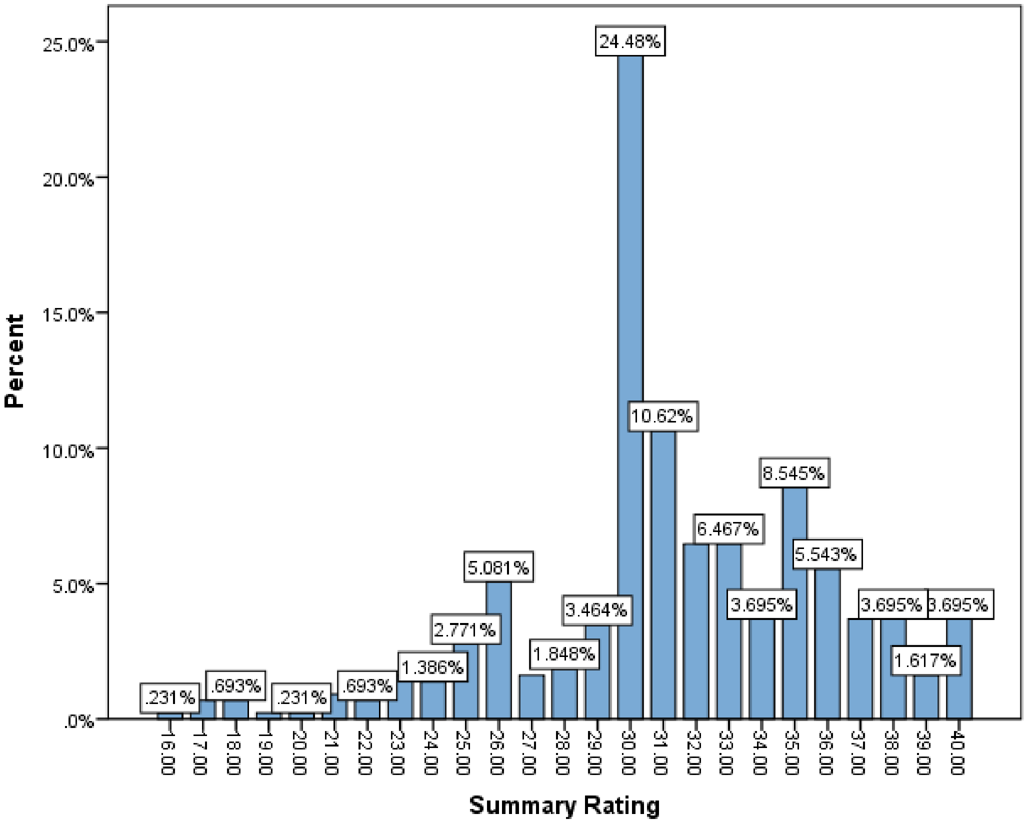


Table 32. Distribution of Translated Summary Ratings

Rating	Frequency	Percent
Unacceptable	8	1.8
Developing/Needs Improvement	33	7.6
Proficient	276	63.7
Exemplary	116	26.8
Total	433	100.0

SIG School Results

Standards 1 to 7 Descriptive Outcomes

Tables 33 to 39 and Figures 9 to 15 show the results of the ratings for Standards 1 to 7 for the teachers from SIG schools. The tables and figures indicate that “Proficient” was the most used category for each standard. Between 47 and 58 percent of the ratings for each standard were in this category. “Exemplary” was the next most often used category with 28 to 46 percent of the teachers being rated in this category per standard. “Developing/Needs Improvement” and “Unacceptable” were not often used for many of the standards with the cumulative percentage rated in those two ranges being generally four percent. The notable exception was Standard 7-Student Academic Progress where 24 percent were rated in those lower categories. Among the standards, Standard 7 had the most variability of ratings.

Table 33: Rating Distribution for Standard 1 – Professional Knowledge

Rating	Frequency	Percent
Unacceptable	1	.3
Developing/Needs Improvement	8	2.4
Proficient	175	51.8
Exemplary	154	45.6
Total	338	100.0

Table 34: Rating Distribution for Standard 2 – Instructional Planning

Rating	Frequency	Percent
Unacceptable	0	0
Developing/Needs Improvement	13	3.8
Proficient	188	55.6
Exemplary	137	40.5
Total	338	100.0

Table 35: Rating Distribution for Standard 3 – Instructional Delivery

Rating	Frequency	Percent
Unacceptable	2	.6
Developing/Needs Improvement	7	2.1
Proficient	195	57.7
Exemplary	134	39.6
Total	338	100.0

Table 36: Rating Distribution for Standard 4 – Assessment of and for Student Learning

Rating	Frequency	Percent
Unacceptable	0	0
Developing/Needs Improvement	10	3.2
Proficient	195	56.9
Exemplary	133	39.9
Total	338	100.0

Table 37: Rating Distribution for Standard 5 – Learning Environment

Rating	Frequency	Percent
Unacceptable	3	.9
Developing/Needs Improvement	16	4.7
Proficient	167	49.4
Exemplary	152	45.0
Total	338	100.0

Table 38: Rating Distribution for Standard 6 - Professionalism

Rating	Frequency	Percent
Unacceptable	1	.3
Developing/Needs Improvement	10	3.0
Proficient	172	50.9
Exemplary	155	45.9
Total	338	100.0

Table 39: Rating Distribution for Standard 7 – Student Academic Progress

Rating	Frequency	Percent
Unacceptable	28	8.3
Developing/Needs Improvement	53	15.7
Proficient	161	47.6
Exemplary	96	28.4
Total	338	100.0

Figure 9: Rating Distribution for Standard 1 – Professional Knowledge

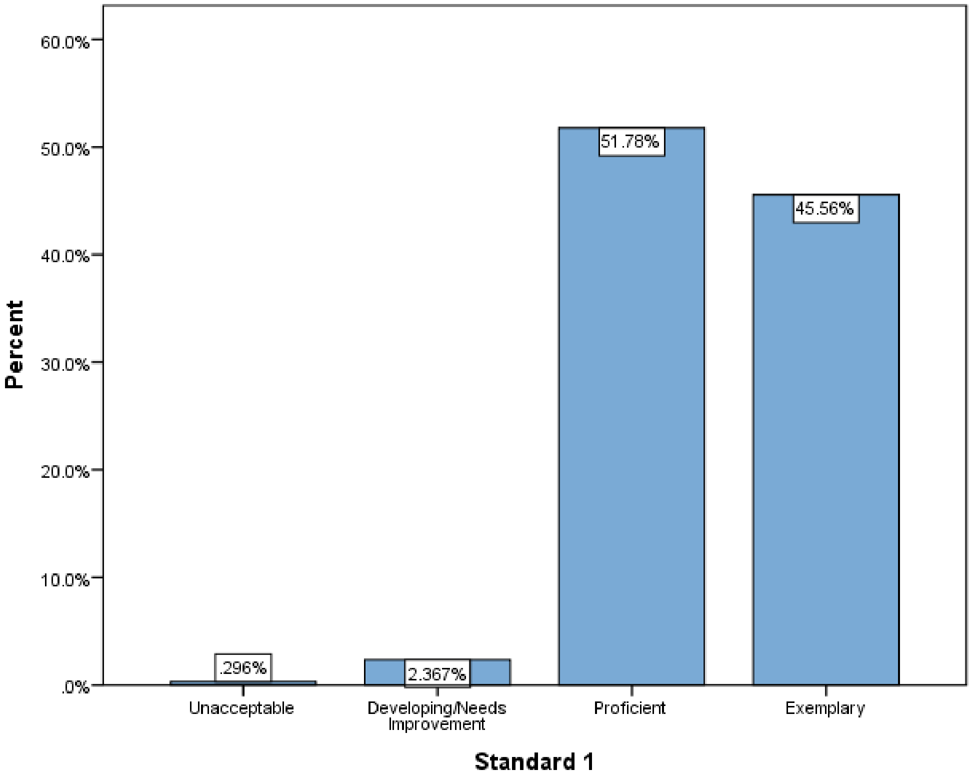


Figure 10: Rating Distribution for Standard 2 – Instructional Planning

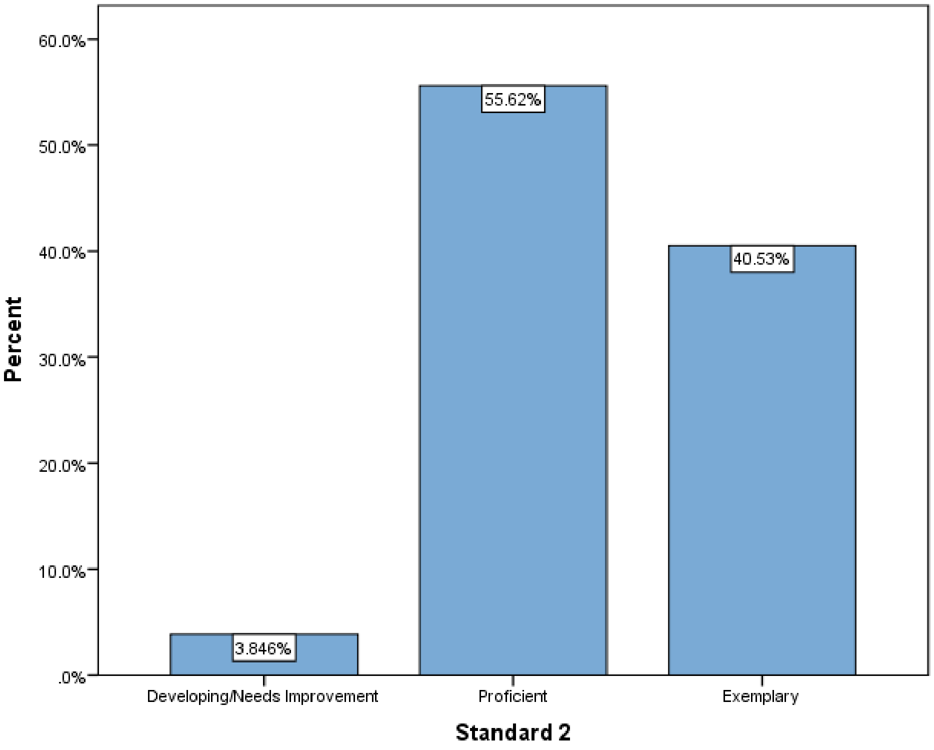


Figure 11: Rating Distribution for Standard 3 – Instructional Delivery

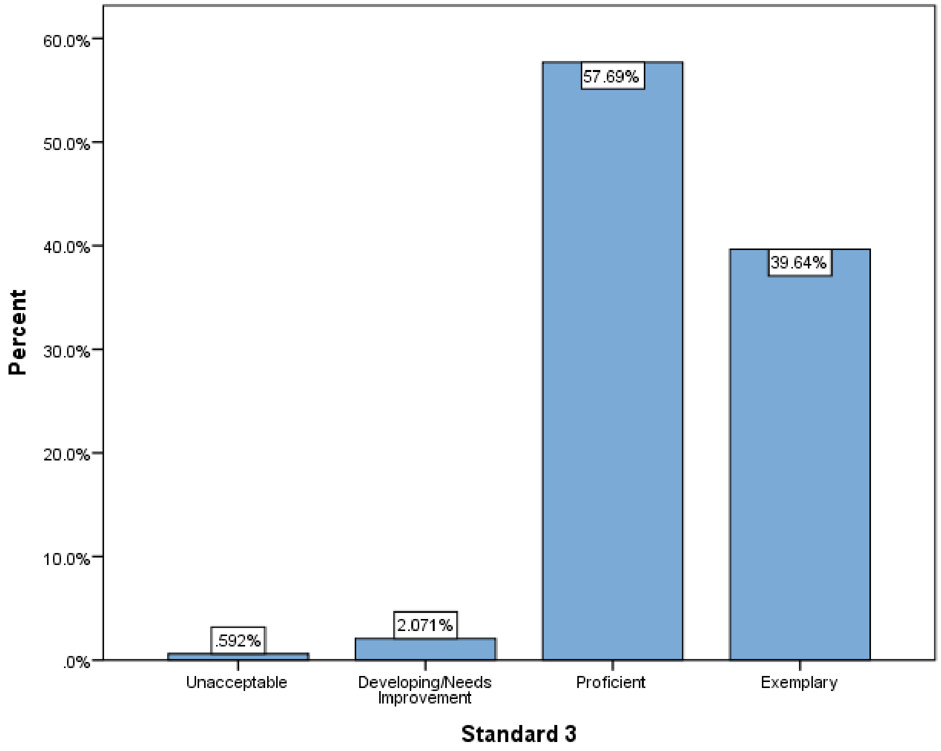


Figure 12: Rating Distribution for Standard 4 – Assessment of and for Student Learning

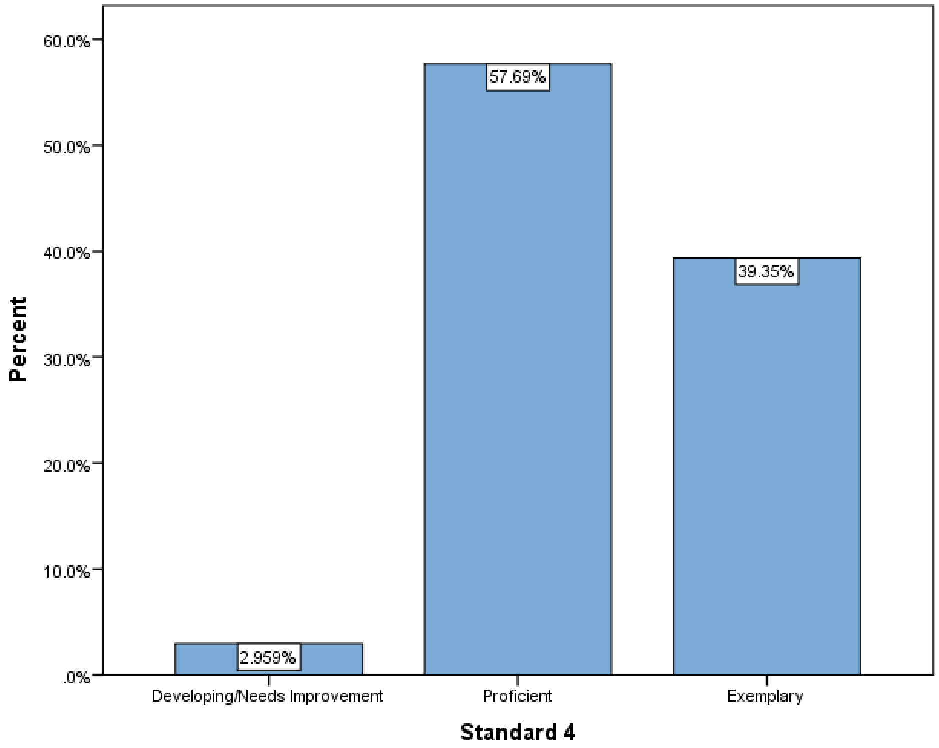


Figure 13: Rating Distribution for Standard 5 – Learning Environment

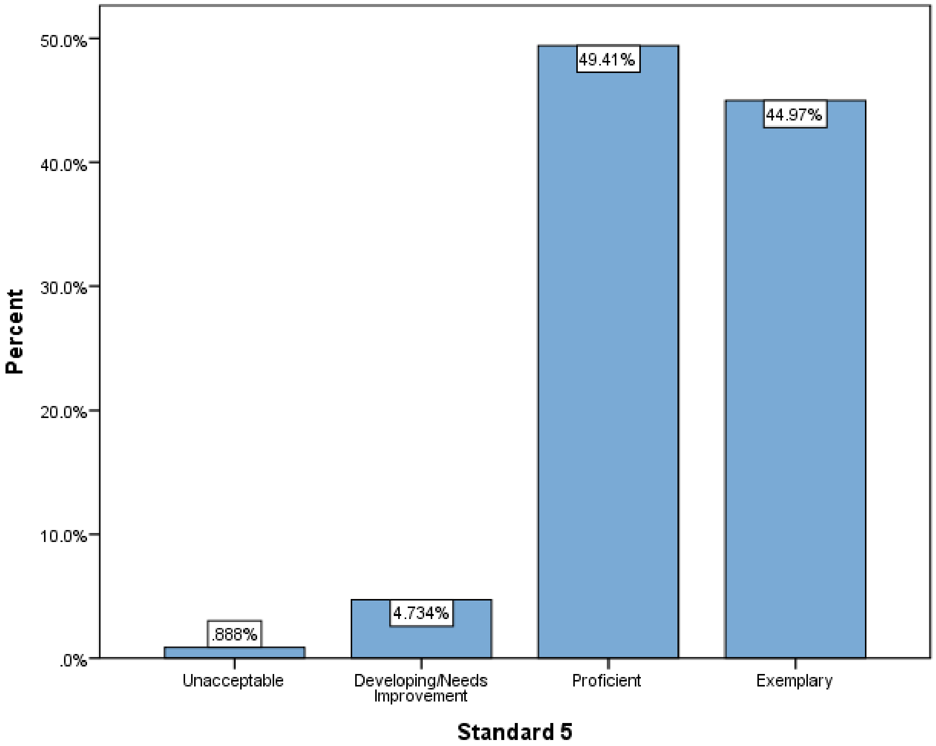


Figure 14: Rating Distribution for Standard 6 – Professionalism

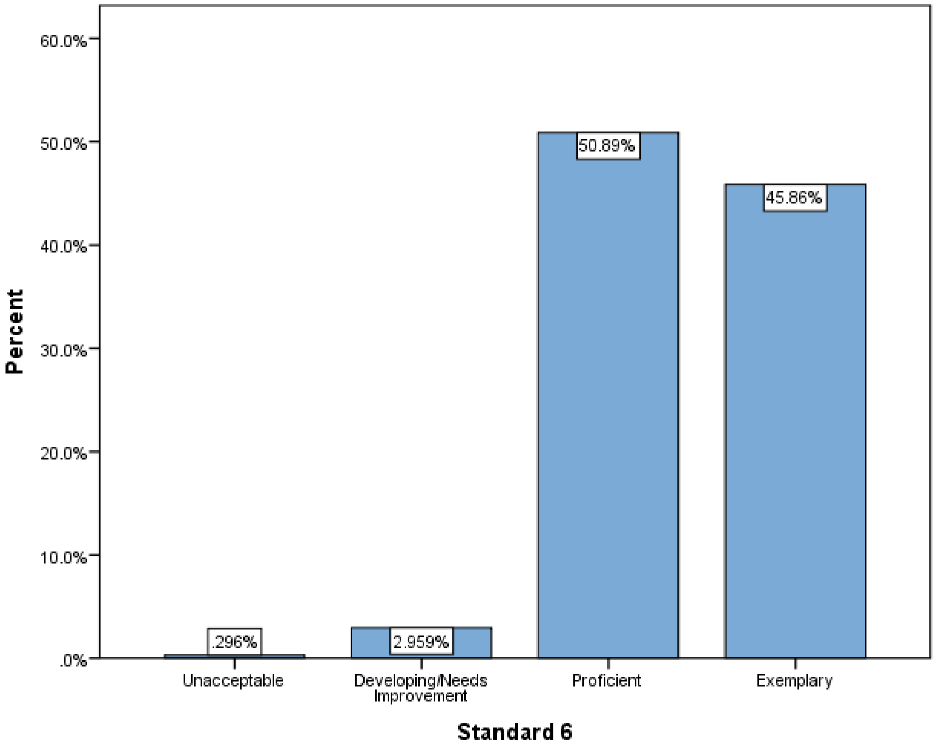
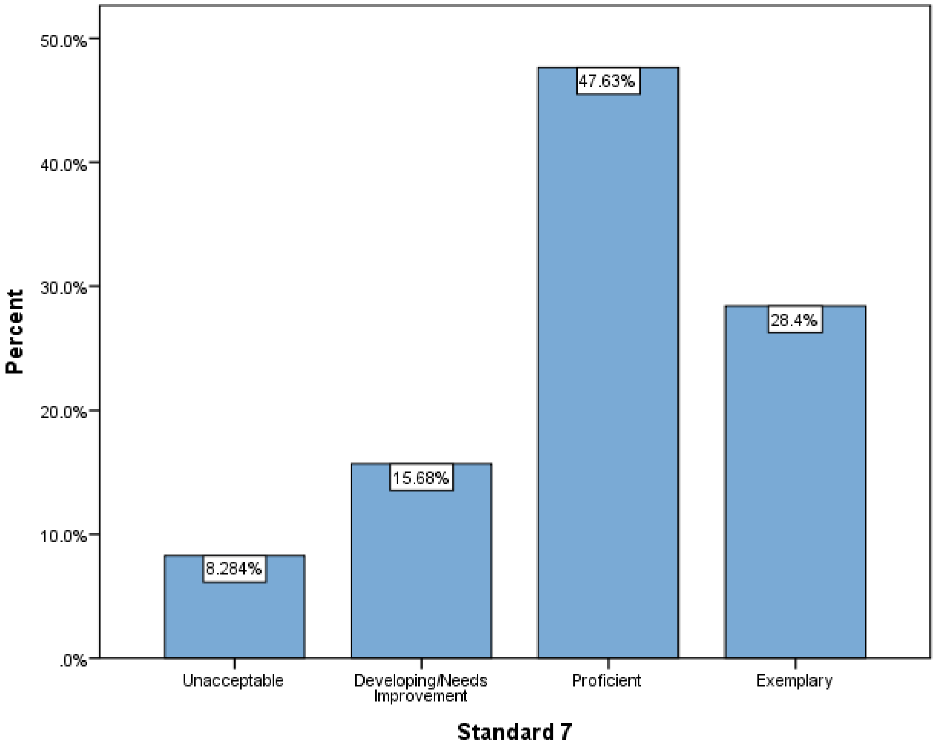


Figure 15: Rating Distribution for Standard 7 – Student Academic Progress



Evaluation Questions

<u>Question 1</u>: What are the relationships among the ratings on the six teacher process standards (Standards 1-6)?

Table 40 provides the correlations among the six process standards (Standard 1 to Standard 6) for the SIG school sample. As can be seen in the table, most of the correlations among the standards are in the moderate range indicating that the ratings of the process standards have considerable overlap. The rating of Standard 1 – Professional Knowledge has the most overlap with other standards. It correlates highly with Standard 3 – Instructional Delivery, Standard 4 – Assessment of and for Student Learning, and Standard 6 – Professionalism. The ratings of Standard 5 – Learning Environment and Standard 6 – Professionalism are the most unique (r = .472). The correlations in the .4 to .5 range indicate that the standards are rated with some common basis. It may be that an overall impression of the teacher's competence underlies all of the ratings and accounts for the correlation among these standards.

Table 40: Correlations Among the Six Process Standards

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and for Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 1: Professional Knowledge	1					
Standard 2: Instructional Planning	.585**	1				
Standard 3: Instructional Delivery	.628**	.535**	1			
Standard 4: Assessment of and for Student Learning	.617**	.623**	.576**	1		
Standard 5: Learning Environment	.563**	.535**	.547**	.519**	1	
Standard 6: Professionalism	.658**	.531**	.517**	.552**	.472**	1

Note: ** correlation is significant at the .01 level. N = 338.

Question 2: To what degree do ratings of teachers' six process standards predict student academic growth as measured by Standard 7 – Student Academic Progress?

Table 41 displays the correlations between the six process standards and Standard 7, which is the rating of student academic progress. The correlations in Table 41 are generally lower than those of Table 40, indicating less overlap between the process standards and student academic progress. All of the correlations are significant and in the moderate range, indicating that there is commonality between all of the process standards and the rating of academic growth. Standard 5 has the weakest relationship to Standard 7 but there is not much practical difference between the correlations indicating a fairly uniform relationship between the process standards and the rating of academic growth.

Table 41: Correlations Between the Six Process Standards and the Achievement Standard

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and For Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 7: Student Academic Progress	.361**	.417**	.326**	.380**	.266**	.378**

Note: ** correlation is significant at the .01 level. N = 338.

The relationship between the process standards and the academic progress measure was expected to be multivariate. That is, it was expected that a combination of process standard ratings would be required to predict the academic outcome rating. To test this, the six process standard ratings were used as predictors in a stepwise multiple regression with Standard 7, the student academic progress rating, as the target. Table 42 presents the results of the regression analysis. The analysis indicated that process standard ratings for Standards 2, 4, and 6 were the only significant predictors of the student academic progress rating, Standard 7. The overall model multiple R was .466 with an R-square of .217. This indicates a modest ability to predict the student academic progress rating from the process standard ratings. The selection of multiple process standard ratings for the model confirms that the student academic progress measure is multivariate in nature. The moderate predictive power of the model indicates that other factors beyond those represented in the six process standards are influential in the student academic progress ratings.

Table 42. Stepwise Regression Results

Summary of Steps

Model	R	R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
a	.417	.174	.171	3.202	.174	70.637	1	.000
b	.456	.208	.203	3.140	.034	14.407	1	.000
c	.466	.217	.210	3.125	.010	4.118	1	.043

a. Predictors: (Constant), Standard 2
b. Predictors: (Constant), Standard 2, Standard 6
c. Predictors: (Constant), Standard 2, Standard 6, Standard 4
d. Dependent Variable: Standard 7

Variable	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.072	1.244		.058	.954
Standard 2	1.519	.407	.241	3.731	.000
Standard 6	1.095	.376	.177	2.915	.004
Standard 4	.866	.427	.133	2.029	.043

Question 3: Are the results for teachers with SGPs similar to the results for teachers who have student achievement goal setting as a major measure of student progress?

The Standard 7 – Student Academic Progress ratings of teachers with and without SGPs were compared to determine whether the availability of SGPs had an impact on the Standard 7 ratings. Table 43 displays the mean Standard 7 ratings (weighted) for the two groups. A t-test (***t*** (df = 334) = .464, p = .643) indicated that there was no difference in the mean ratings of the two groups. This finding suggests that there was no differential influence of SGPs availability on the Standard 7 ratings. This finding can be considered encouraging in that it suggests that evaluators were able to apply both Student Achievement Goal Setting data and Student Growth Percentile data in a manner that yielded comparable ratings for Standard 7.

Table 43. Standard 7 Descriptive Statistics for Teachers With and Without SGPs

SGPs Used	N	Mean	Standard Deviation	Standard Error Mean
No	310	11.82	3.59	.20
Yes	26	12.15	2.65	.52

Although the evaluation question did not directly ask about differences in the process standard ratings, they were tested to see whether differences existed between the SGPs and non-SGPs groups. Table 44 shows the descriptive statistics and the t-test values for the six process standards by SGPs group. There were no significant differences on the six standards. This indicates that teachers who have SGPs data available were perceived by evaluators in a similar manner to those without SGPs data.

Table 44. Standards 1 through 6 Descriptive Statistics for Teachers With and Without Student Growth Percentiles (SGPs)

Standard	SGPs Used?	N	Mean	Std. Deviation	Std. Error Mean	t value
Standard 1	No	310	3.41	.56	.03	-1.44
	Yes	26	3.58	.50	.10	
Standard 2	No	310	3.35	.56	.03	-1.61
	Yes	26	3.54	.51	.10	
Standard 3	No	310	3.35	.56	.03	-1.68
	Yes	26	3.54	.51	.10	
Standard 4	No	310	3.35	.54	.03	-1.32
	Yes	26	3.50	.51	.10	
Standard 5	No	310	3.39	.63	.04	0.37
	Yes	26	3.35	.49	.10	
Standard 6	No	310	3.41	.57	.03	-1.78
	Yes	26	3.62	.50	.10	

** Groups are significantly different at p < .01 level.

Note: The ratings of the Standards have been treated here as interval type indicators. The ratings of the standards could also be considered as ordinal indicators. The statistical treatment of such indicators would be different. Analyses were also conducted treating the standard ratings as ordinal data with the results indicating the same conclusions as when the standard ratings were treated as interval data.

Question 4: Do school principals sufficiently discriminate in the application of the teacher evaluation system based on measures of effectiveness?

Consideration of this question requires an examination of the summary ratings. To arrive at the summary rating, raters were instructed to combine the six process standards ratings with the student academic progress standard (Standard 7). Standard 7 was to be weighted 40 percent and each of the other standards was to be weighted as ten percent. Examination of the reported data indicated that not all summary ratings conformed to this structure. For instance, some summary ratings were on a scale of 1 to 4, while some were in the form of text (e.g., “Exemplary,” “Proficient,” “Developing/Needs Improvement,” or “Unacceptable”). Therefore, the summary ratings were recalculated based on the approved formula provided in the *Guidelines* to ensure the data were in a consistent format while not changing the principals’ actual ratings. The formula used was: Summary Rating = (Standard 1 Rating × 1) + (Standard 2 Rating × 1) + (Standard 3 Rating × 1) + (Standard 4 Rating × 1) + (Standard 5 Rating × 1) + (Standard 6 Rating × 1) + (Standard 7 Rating × 4). This process did not alter the distribution and the variability of the distribution of teachers’ evaluation results within the school but did enable the examination of rating across schools.

Table 45 shows the descriptive statistics for the summary ratings. Summary ratings could vary between 16 and 40. The table shows that the mean value was toward the higher end of the rating scale with a standard deviation of approximately 5 points. No teachers were rated below 16 but some were rated at the top of the scale. Figure 16 is a graphical representation of the summary ratings. This figure confirms that summary ratings cluster on the higher side of the modal point 30. The figure also confirms that there is variability among the summary ratings but that the lower end of the possible distribution is not utilized as often as the top end.

Summary ratings can be translated into the descriptive categories of the original standards. Values 10 through 19 indicate “Unacceptable”; 20 through 25 indicate “Developing/Needs Improvement”; 26 through 34 indicate “Proficient”; 35 through 40 indicate “Exemplary”. Table 46 shows the distribution of summary ratings by category. Based on these percentages, 39 percent of the teachers were rated in the “Exemplary” range and approximately ten percent were rated “Developing/Needs Improvement” or “Unacceptable”. Again, this table indicates that discrimination is occurring but mostly at the upper range of the rating scale.

Table 45. Descriptive Statistics for the Summary Rating.

	Range	Minimum	Maximum	Mean	Std. Deviation
Summary Rating	24.00	16.00	40.00	32.17	5.30

Figure 16. Summary Ratings

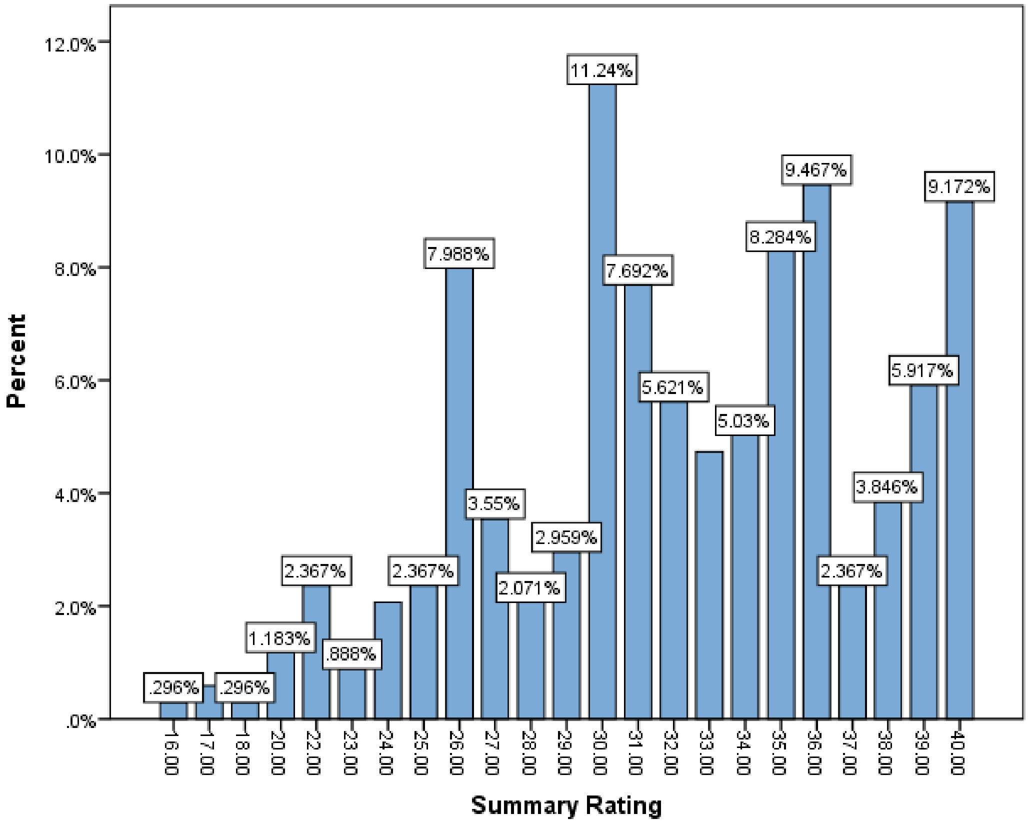


Table 46. Distribution of Translated Summary Ratings

Rating	Frequency	Percent
Unacceptable	4	1.2
Developing/Needs Improvement	30	8.9
Proficient	172	50.9
Exemplary	132	39.1
Total	338	100.0

Lessons Learned and Recommendations

This report examines the validity of the 2011-2012 Virginia Teacher Performance-Pay Initiatives (VPPI) Pilot. During the pilot, administrators, key instructional leaders, and teachers from participating schools were provided with extensive support to restructure their total teacher evaluation systems in alignment with state guidelines and policy, including rigorous teacher performance standards, student achievement growth measures, standardized evaluation protocols, and linking results to teacher performance pay. Evaluators in the pilot schools were trained to use rating rubrics to make summative ratings of the seven standards of each participating teacher. The rating rubrics described four levels of how well the standards were performed on a continuum from "Exemplary" to "Unacceptable." The use of the scale enabled evaluators to acknowledge teachers who exceeded expectations (i.e., "Exemplary"), note those who met the standard (i.e., "Proficient"), and use the two lower levels of feedback for teachers who did not meet expectations (i.e., "Developing/Needs Improvement" and "Unacceptable").

Participating teachers received ratings on six process standards and one outcome standard – Standard 7 – Student Academic Progress. Student academic progress was measured by Student Growth Percentiles (where available and appropriate), student achievement goal setting, and other relevant measures. This study examined the internal validity of the system, and specifically, found:

1) The ratings on the six process standards were positively related to each other, and the correlations among them were consistent.
2) There was a significant correlation between each of the six process standards and student academic growth. However, the stepwise multiple regression analysis indicates only three process standard ratings were the significant predictors of the Standard 7 ratings. It also revealed that the process standards had a modest ability to predict academic outcomes. That is, teachers with high ratings on process standards tend to generate high student academic growth, and student growth decreases as the teachers' ratings on process standards decrease. Nevertheless, the moderate predictive power indicates that other factors that have not been captured in the six process standards are influential in student academic progress ratings.
3) The availability of SGPs had no significant impact on the Standard 7- Student Academic Progress ratings. The ratings on Standard 7 – Student Academic Progress for teachers with SGPs were comparable to the ratings for teachers who had student achievement goal setting as a major measure of student progress.
4) The new teacher evaluation system allowed principals to discriminate their ratings of teachers' performance based on measures of their effectiveness, with approximately 32 percent of teachers (248 out of the 771 teachers included in the data analysis) rated as "Exemplary", 10 percent (75 of 771) rated as "Developing/Needs Improvement" or "Unacceptable," and the majority (58 percent, 448 of 771) rated as "Proficient." Discrimination was occurring but mostly at the upper range of the rating scale. This result was expected, since the rating of "Proficient" was the expected level of performance, and this should be what most teachers were rated. However, there appears to be a higher-than-expected percentage of teachers receiving "Exemplary" ratings in all six of the process standards.

Overall, a moderate relationship between student achievement growth measures and teacher evaluation scores based on performance standards was established. This finding implies the new evaluation system was measuring teacher effectiveness validly to a substantial degree. The relationship was possibly mediated by factors that have not been examined, for instance, the fidelity of implementation. The following recommendations are provided to guide the further implementation of the evaluation system on a statewide basis:

- Conduct further research to improve the validity of the evaluation system design;
- Provide more extensive training with in-depth follow-up to enhance fidelity of implementation. Thorough, quality training is essential to ensuring that school leaders have the tools and skills they need to confidently evaluate and give feedback to teachers;
- Establish clear and consistent communication, which is critical to help teachers and other stakeholders feel confident in the implementation of the new evaluation system; and
- Make knowledge of new developments in teacher evaluation part of leadership and teacher preparation programs.

The results presented in this study are still preliminary. Additionally, it should be noted that the 25 pilot schools likely are not representative of all Virginia schools, in particular because the 25 were selected based on their historical low student performance or their qualification as hard-to-staff. Further, it takes time and practice for new evaluation systems, or any reform policies, to be institutionalized in schools. However, the 2011-2012 pilot study does show that inflation in ratings is happening, with approximately 90 percent of participating teachers being rated as proficient or higher. Ongoing, rigorous, and refresher training sessions should be provided to ensure that evaluators will be able differentiate effective from ineffective teachers consistently over time. Support, such as creating libraries of videos, should be sustained and accessible so that raters can refresh their memories of the examples of teachers' performance along the continuum of "Exemplary" to "Unacceptable." In addition, large-scale research studies need be conducted as the system is being implemented statewide to continuously monitor the alignment among the different measures.

Endnotes

[1] Stronge, J. H. (Ed.). (2006). *Evaluating teaching: A guide to current thinking and best practice* (2nd ed.). Thousand Oaks, CA: Corwin Press. p. 1.

[2] Barber, M. & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. London: McKinsey & Company. Retrieved from http://www.mckinsey.com /locations/ukireland/publications/pdf/Education_report.pdf

[3] Tucker, P. D., Stronge, J. H., & Gareis, C. R. (2002). *Handbook on Teacher Portfolios for Evaluation and Professional Development.* Larchmont, NY: Eye on Education.

[4] Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79-107.

[5] Westberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from www.widgeteffect.org

[6] Stronge, J. H., & Tucker, P. D. (2003). *Handbook on teacher evaluation: Assessing and improving performance.* Larchmont, NY: Eye on Education.